Study of Variability in Advanced Transistor Technologies

By

Nattapol Damrongplasit

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

and the Designated Emphasis

in

Nanoscale Science and Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Tsu-Jae King Liu, Chair
Professor Sayeef Salahuddin
Professor Junqiao Wu

Fall 2014

Study of Variability in Advanced Transistor Technologies

by

Nattapol Damrongplasit

Abstract

Study of Variability in Advanced Transistor Technologies

by

Nattapol Damrongplasit

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

As transistor dimensions are scaled down in accordance with Moore's Law to provide for improved performance and cost per function, variability in transistor performance grows in significance and can present a major challenge for achieving high yield in the manufacture of integrated circuits utilizing transistors with sub-30 nm gate lengths. Increased variability in the threshold voltage (V_T) of a transistor ultimately limits the minimum operating voltage for six-transistor (6T) static memory (SRAM) cells, hinders aggressive scaling of cell area, and causes performance degradation in analog circuits. Better understanding and accurate assessment of device variation are needed in order to minimize yield loss and design margin.

Several variability reduction techniques and variability characterization/modeling methodologies are explored in this work. Device simulations are performed to assess the benefit of super-steep retrograde (SSR) channel doping to reduce variability in transistor performance and thereby extend the scalability of planar bulk-Si CMOS technology with minimal incremental cost. Variability analysis of a 32nm high-permittivity-dielectric/metal gate (HKMG) stack CMOS technology using current-*vs*.-voltage characteristics of transistors operated in forward (F) and reverse (R) modes measurements is used to explain variabilities in V_T and in drain-induced barrier lowering (DIBL) and their correlations, which cannot be captured by a traditional SPICE modeling. Test chips are designed for characterization of systematic and random variability in 16nm and 28nm generation Fully Depleted Silicon-On-Insulator (FDSOI) technologies via device arrays and padded-out SRAM cells.

The effect of random variability on the performance of a tunnel-field effect transistor (TFET) is also examined. The TFET has emerged as a promising candidate to replace the MOSFET for low-power applications, due to its promise of achieving higher I_{ON}/I_{OFF} at low operating voltages. Three-dimensional (3D) device simulations are used to simulate the effects of random dopant fluctuations and line edge roughness on the performance of planar Ge-source and a raised-Ge-source TFET structures.

To my family,
for their unbounded love and unwavering support

Table of Contents

Chapter 4: Variability Characterization in Fully-Depleted Silicon-On-Insulator (FD-SOI) Transistors

Acknowledgements

######

Chapter 1

Introduction

1.1 Transistor, Circuit, and Moore's Law: A Historical Perspective

During the first half of the twentieth century, electronic circuits were powered by vacuum tubes which are bulky, expensive, power-hungry, and notoriously unreliable [1]. The state-of-the-art computing machine ENIAC (Electronic Numerical Integrator And Computer) used for calculating missile trajectories was made out of 18,000 vacuum tubes, and it occupied a large room [2]. To overcome these short-comings of the vacuum tube, much research was aimed on finding its replacement. In 1947, three scientists (John Bardeen, Walter Brattain, and William Shockley) at Bell Laboratories were credited for inventing what is now considered as the first transistor [2]. Compared to a vacuum tube, a transistor is much more power-efficient and far more reliable. Fast forward about ten years after, Jack Kilby at Texas Instruments introduced the world to the first integrated circuit (1958), where two transistors were connected together to build a flip-flop [1]. As simplistic as that circuit was, it was a monumental point for the age of digital electronics since it provided a glimpse into the potential of integrating multiple transistors on the same chip, having them work together to perform a more complex operation. In 1965, Gordon Moore who was working at Fairchild Corporation and later co-founded Intel Corporation made an observation that the number of transistors on a single die doubled every 18 months, and he predicted that this exponential growth would continue (Fig. 1.1a) [1,3]. This prediction has proven to be surprisingly accurate and it later became known as *Moore's Law* [4-5]. It is arguably one of the most important laws which helped to usher in the digital era. Not only did Moore's original paper state that integrating more transistors would increase performance, but it also highlighted the fact that the cost of manufacturing each transistor can be cheaper as shown in Fig. 1.1b [3]. This has driven the semiconductor industry towards increasing transistor density on a chip ever since.

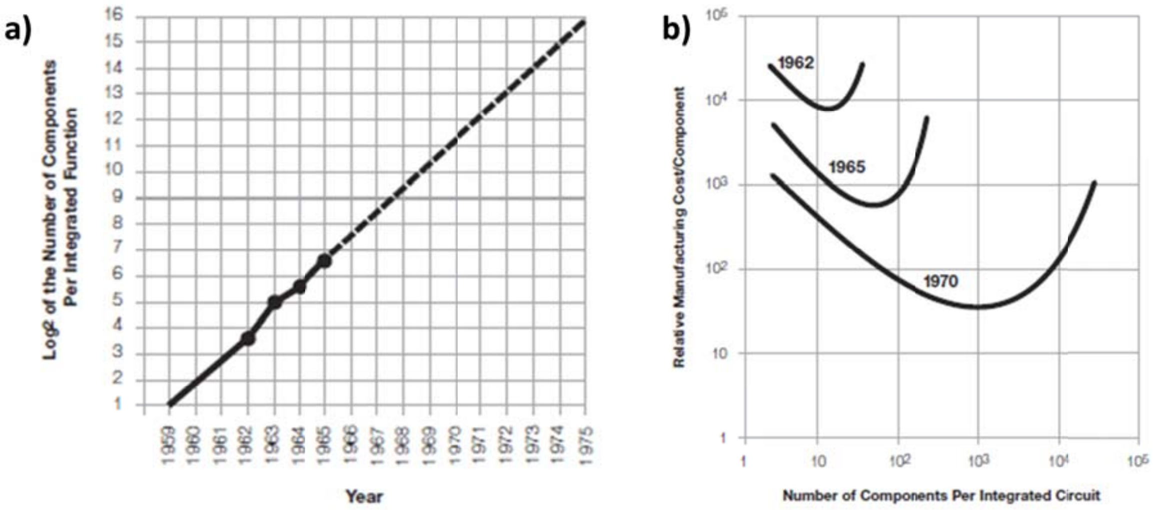
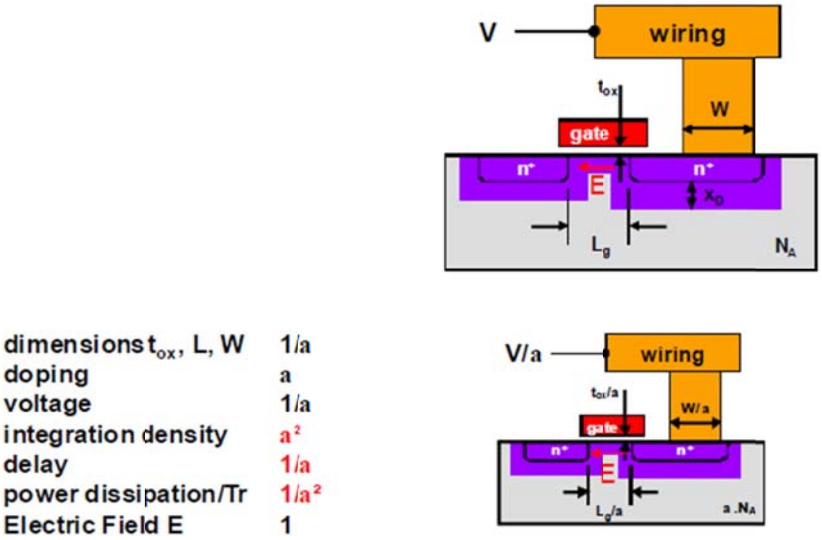


Fig. 1.1 a) Moore's original paper showing number of transistors in log-scale vs. year in production b) Plot of cost vs. the number of transistors per integrated circuit. There exists an optimum point where the cost of manufacturing is lowest (adapted from [3]).

1.2 Challenges in Transistor scaling

To achieve higher transistor density on a chip, the dimensions of a transistor must be scaled, so that more can be packed into a given die area. In 1974, Robert Dennard at IBM TJ Watson Research Center proposed a way to scale a transistor while preserving its basic operation and also reaping the performance benefit that comes with scaling [6]. His approach is known as constant field scaling, where device dimensions, doping density, and supply voltages are scaled simultaneously (Fig. 1.2) [7]. The approach worked for a while but started to slow down dramatically when the minimum half-pitch reached 90 nm. As the gate length of a transistor is made smaller, non-ideal effects, which are negligible at long gate lengths, can degrade transistor performance. These effects are usually referred to as short-channel effects (SCEs)[8]. For example, as the gate length is scaled down, the influence of the applied drain voltage becomes so strong such that a significant amount of current can still flow even when a transistor is supposed to be OFF. The thickness of the gate dielectric also needs to be thinned down in order to maintain good capacitive coupling between the gate and the channel potential. However, as the physical thickness of the gate dielectric approaches 1 or 2 nm, quantum mechanical tunneling of carriers directly from the channel through the gate dielectric results in undesirable gate leakage current. Ideally, the transistor operating voltage should be reduced along with the gate length. As shown in Fig. 1.3, reductions in the supply voltage (V_{DD}) of complementary metal-oxide-semiconductor (CMOS) digital logic circuits have slowed below 1V [9]. The main reason for this is that the threshold voltage (V_T) of an MOS transistor cannot be scaled down aggressively. As V_T is lowered, the OFF-state leakage current and hence static power dissipation increases exponentially (Fig. 1.4). Since V_T cannot be scaled down easily, V_{DD} also cannot be scaled down much below 1 V because sufficient gate overdrive (V_{DD}-V_T) is required for a transistor to have sufficient ON-state current to meet circuit operating speed requirements.



dimensions t_{ox}, L, W	1/a
doping	a
voltage	1/a
integration density	a^2
delay	1/a
power dissipation/Tr	$1/a^2$
Electric Field E	1

Fig. 1.2 Constant field scaling of a MOS transistor wherein device dimensions, doping, and voltage are scaled proportionally, resulting in higher integration density, lower circuit operating delay, and less power dissipation per transistor [7].

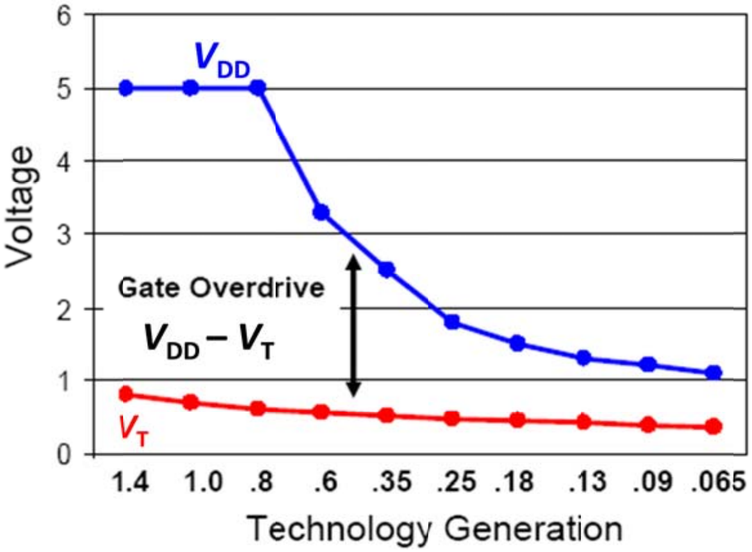


Fig. 1.3 Historical scaling of CMOS supply voltage (V_{DD}) and threshold voltage (V_T) vs. technology node. V_{DD} scaling has slowed down significant after 90 nm node (adapted from [9]).

.

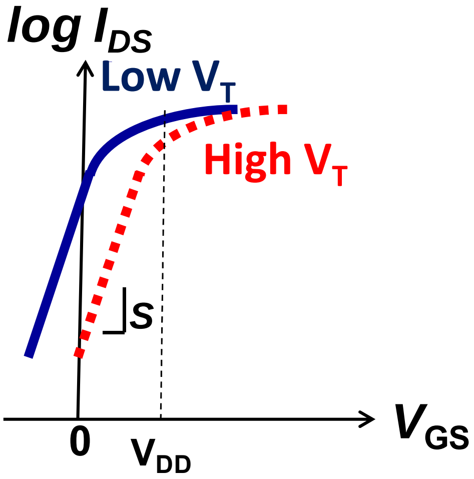


Fig. 1.4 Transfer characteristic I_{DS}-V_{GS} of a typical N-channel MOSFET, comparing between low- and high-V_T transistors. A linear decrease in V_T results in an exponential increase in OFF-state leakage current at $V_{GS} = 0$ V.

The energy consumed per operation of a CMOS digital circuit can be broken down into dynamic and static components as shown in Eqn. 1.1 – 1.2 and graphically plotted in Fig. 1.5.

$$E_{total} = E_{dynamic} + E_{static} = \alpha L_d C V_{DD}^2 + L_d I_{OFF} V_{DD} T_{delay} \tag{1.1}$$

$$T_{delay} = \frac{L_d C V_{DD}}{2 \cdot I_{ON}} \tag{1.2}$$

where α = switching activity factor, L_d = logical depth, C = gate capacitance per stage, V_{DD} = supply voltage, I_{OFF} = OFF-state current, I_{ON} = ON-state current.

For higher computational throughput, transistors are pushed to run at faster operating frequency as more of them are integrated on a chip. Increasing frequency without scaling down V_{DD} results in higher dynamic power consumption. On top of this, as V_T has been lowered, transistors have become more leaky resulting in higher static power consumption (Fig. 1.6) [10]. All of this led to a CMOS power crisis, where the power density of a chip was increasing at an alarming rate. This prompted the industry to adopt parallelism (with the introduction of multiple cores rather than increasing the clock speed of a single core) as the primary means of improving throughput within a power constraint, as shown in Fig. 1.7. The overall performance of the integrated circuit chip can still be improved by having multiple slower cores running in parallel.

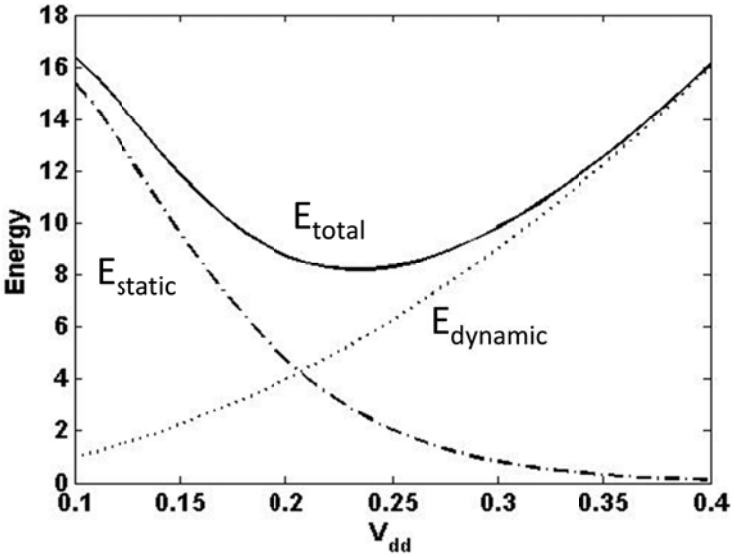


Fig. 1.5 Energy consumed per digital operation of a CMOS circuit is comprised of dynamic and static components. There exists a lower limit for the total energy per operation.

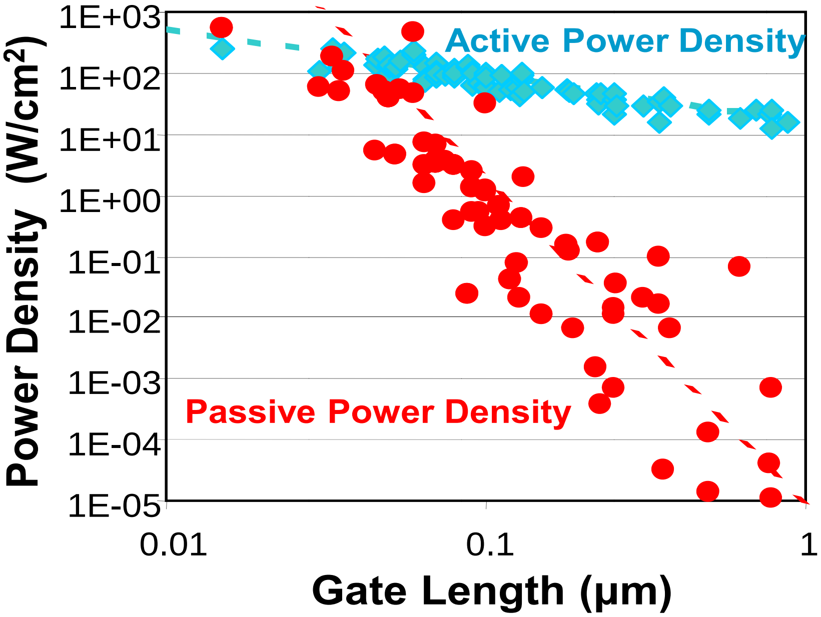


Fig. 1.6 Power density vs. transistor gate length. At smaller gate lengths, static (passive) power density becomes comparable to dynamic (active) power density. (adapted from [10]).

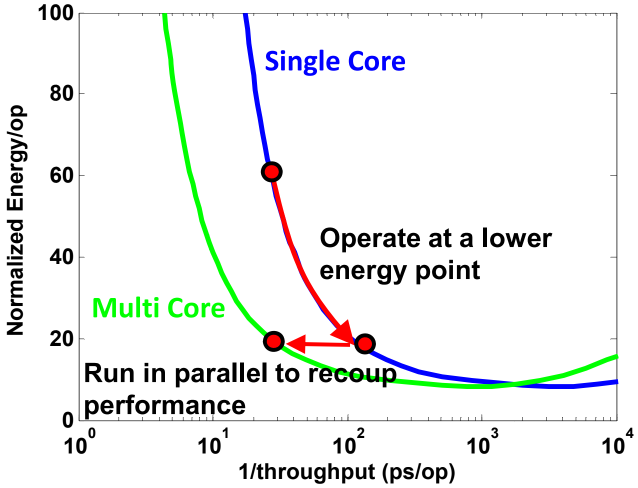
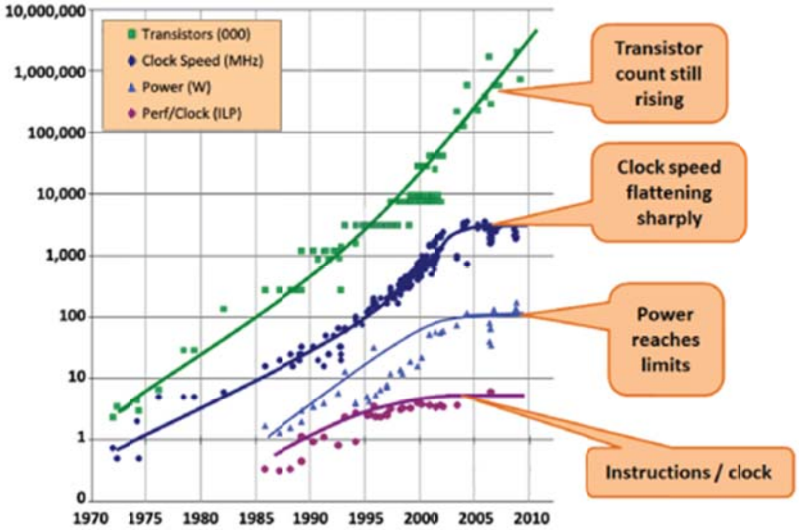


Fig. 1.7 Energy per operation vs. delay. Performance improvement can be achieved by running multiple cores at slower speed (longer delay), in parallel.



Source: Intel

Fig. 1.8 Trends in transistor counts, clock speed, power dissipation, and number of instructions. Note that the increase in clock speed has flattened over the years due to limits for chip power density.

1.3 Advanced Transistor Structures

Short-channel effects limit the performance benefits that come with the ‘classical’ (Dennard) scaling of the MOS field-effect transistor (FET). To compensate for this diminished return, the MOSFET structure has evolved over the years to mitigate these effects.

1.3.1 Planar Bulk MOSFET

The most widely used transistor design today is still the classical planar bulk MOSFET [11]. It was continually miniaturized over three decades to achieve improvements in chip performance and cost per function. Fig. 1.9 shows a series of transmission electron microscopy (TEM) images of transistors from the 90nm technology node down to the 32nm technology node [12]. To help boost the ON-state drive current, channel strain engineering has been used to increase carrier mobility [13]. For n-channel (NMOS) transistors, the stress memorization technique (SMT) and a tensile contact-etch-stop-liner (CESL) are often used to induce tensile stress along the channel direction, which is beneficial for electron mobility [14-15]. For p-channel (PMOS) transistors, embedded silicon-germanium (SiGe) source/drain regions are the primary method of generating compressive stress along the channel direction, which is beneficial to hole mobility. Gate leakage due to the thin layer of gate oxide (SiO_2) can be mitigated by using a high-permittivity (high-k) dielectric and metal gate (HKMG) stack [16]. By using a high-k dielectric rather than SiO_2, it is possible to have a physically thicker insulating layer while keeping the same capacitive coupling, resulting in reduced quantum mechanical tunneling and hence lower gate leakage current. Additionally, another effective way to suppress SCE is to engineer the channel doping profile. Techniques such as halo and retrograde well doping are widely used to help improve electrostatic control of the channel and suppress OFF-state leakage current [18-19].

Short-channel effects become very difficult to suppress in the planar bulk transistor for gate lengths below 25nm. Therefore, to allow further miniaturization, advanced transistor structures employing thin-body regions such as the fully depleted silicon-on-insulator (FD-SOI) and three-dimensional FinFET have been developed for future generations of CMOS technology [20-21].

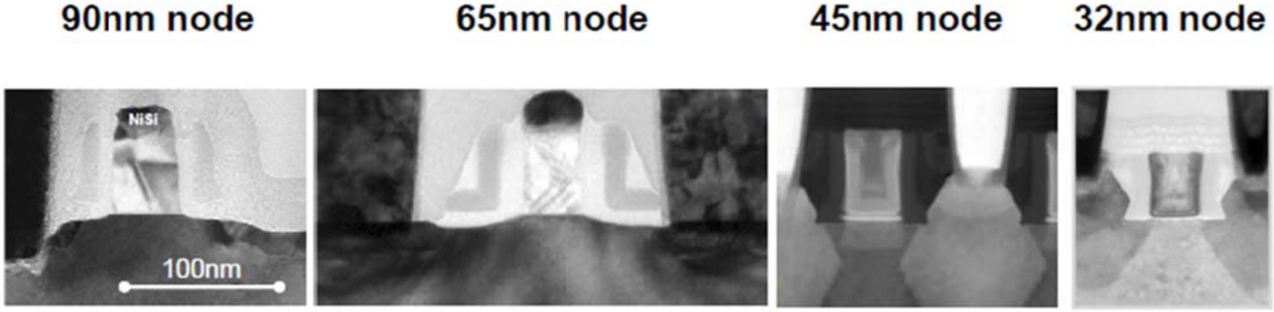


Fig. 1.9 TEM cross-sections of MOSFETs at various technology nodes, with approximately the same scale.

1.3.2 FD-SOI (Fully-Depleted Silicon-On-Insulator) MOSFET

Similar to a conventional bulk transistor, an FD-SOI MOSFET is a planar structure. It is fabricated in a thin Si layer on top of a buried oxide (BOX) layer. An illustration of the FD-SOI MOSFET structure is shown in Fig. 1.10a and a TEM cross-section is shown in Fig. 1.10b [22-23]. Electrostatic gate control in the FD-SOI MOSFET is superior to that in a planar bulk MOFSET thanks to the thin silicon body (since OFF-state leakage current paths far away from the gate are eliminated) [23]. With a very thin body, channel doping is not necessary to suppress SCE and thus can be low. The scalability of the FD-SOI MOSFET structure is improved if the thickness of the BOX layer is also scaled down and "ground plane" doping underneath the BOX layer is employed [24-25]. Back-biasing is also possible, allowing for dynamic V_T tuning. Because of the close similarity between FD-SOI and planar bulk MOSFETs, minimal changes are needed to porta circuit design from bulk to FD-SOI.

Although the FD-SOI MOSFET is promising to replace the planar bulk MOSFET at gate lengths below 25 nm, it does have limitations. In order to maintain good electrostatic integrity, the silicon thickness T_{si} should be at least 3 times as small as the gate length ($T_{si} \sim L_g/3$) [26]. This small thickness requirement makes FD-SOI transistor more susceptible to T_{si} variation. In terms of mobility enhancement, it is found that SiGe S/D and gate stressors are not very effective in transferring the stress into the channel [12]. Additionally, the starting SOI wafer has to be made by a special process, which adds to the cost of manufacturing and can limit the overall supply chain.

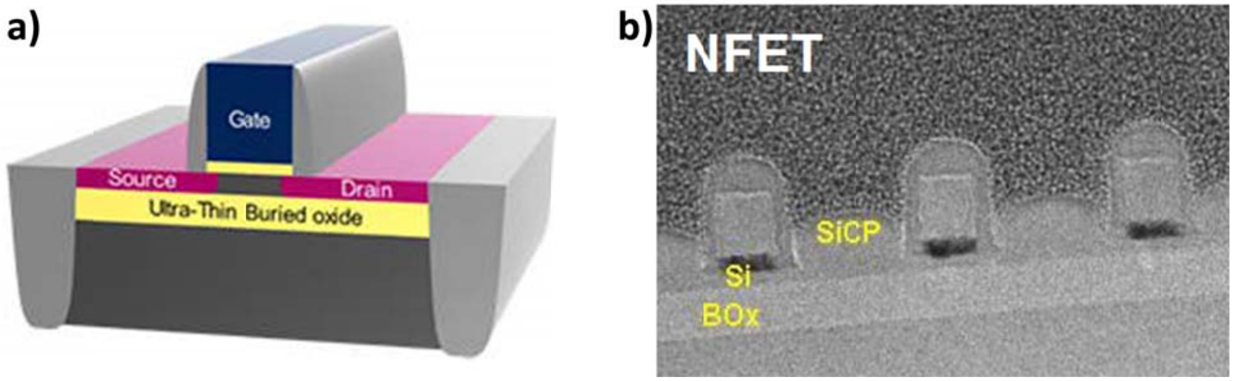


Fig. 1.10 a) A 3D view of a planar FD-SOI transistor [22]. b) A TEM cross-section of N-channel FD-SOI transistors [23].

1.3.3 FinFET

Another variation of a thin-body MOSFET is the vertical FinFET, where the body has a fin-like shape and a gate electrode straddles it, as depicted in Fig. 1.11a [20,27-30]. It can be made on a bulk silicon or SOI substrate [31]. Due to gating from all 3 sides of the channel, SCE can be well suppressed if the fin width is less than 1/2 the gate length ($W_{si} \sim L_g/2$). Experimental results have demonstrated FinFET transfer characteristics with low subthreshold swing ~70 mV/dec and low drain-induced barrier lowering (DIBL) ~50 mV/V [20]. Additionally, since the effective transverse electric field in the inversion-layer channel is lower and the efficiency of stress transfer to the channel region is more than 50% for various stressors, field-effect mobility is substantially enhanced [12, 32]. Recently, the FinFET has supplanted the planar bulk MOSFET in the most advanced microprocessor chips produced by Intel Corporation, and the leading semiconductor foundries plan to introduce FinFETs in future CMOS production processes.

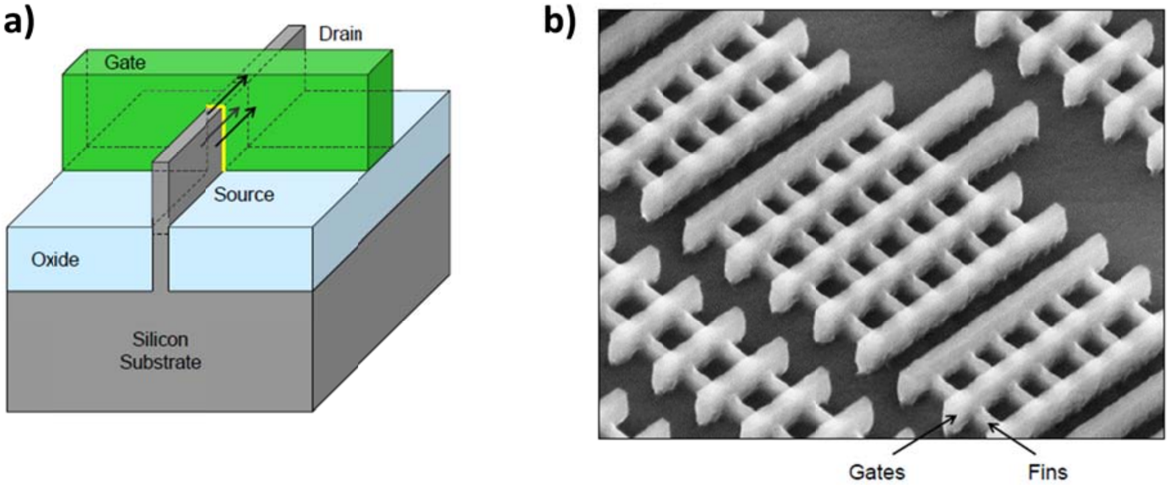


Fig. 1.11 a) A 3D view of a vertical FinFET transistor b) TEM image of an array of FinFET transistors showing the fin and gate features. [30]

Unlike planar transistors, the FinFET is a "three-dimensional" ("3D") structure, which presents additional challenges for manufacturing and circuit design. For example, ensuring good electrostatic control of a FinFET transistor requires that the fin width be no more than one-half the gate length (i.e. $W_{si} \leq L_g/2$) [26, 32]. Although this thinness requirement is less stringent, compared to that for T_{si} in a FD-SOI MOSFET, it is very challenging to fabricate a high-aspect-ratio structure with good control for acceptable manufacturing yield. Since the width and the height of a FinFET is usually fixed for a given chip, the drive current of a FinFETis tuned by changing the number of fins, leading to current discretization which places restrictions on design [33]. The series resistance of the source/drain regions in a narrow fin can limit transistor performance, especially for analog and RF circuit applications [34]. For System-on-Chip (SOC) applications, different values of V_T are not straightforward to achieve due to the lack of a body-biasing effect. Thus, V_T tuning will likely have to rely on gate work function turning and gate-length trimming instead. The 3D nature of a FinFET also results in larger parasitic gate capacitance as compared to a planar transistor [32].

1.3.4 Tunneling FET (TFET)

Advanced structures such as the FD-SOI MOSFET and FinFET can certainly help to improve the performance of a transistor and allows for further technology scaling. Nevertheless, CMOS has a fundamental limit on energy per operation as shown previously in Fig. 1.5. Governed by Boltzmann statistics, the theoretical minimum subthreshold swing (SS) of a MOSFETis 60 mV/dec at room temperature. Any prospective MOSFET-replacement device should have a SS that is smaller than this limit in order to provide for lower energy computing. For the same leakage current, this new device with a smaller SS will be able to achieve a higher on-state drive current (Fig. 1.12a), resulting in a lower energy per operation for a given operating frequency (Fig. 1.12b). There are currently a number of device candidates that have the potential for a very steeply switching operation; these include tunneling FETs, the negative capacitance MOSFET, and nano-electro-mechanical (NEM) relays [34-39]. The TFET in particular has emerged as a strong MOSFET-replacement candidate due to its close similarity to a conventional MOSFET. Fig. 1.13 shows a basic comparison between the MOSFET and tunnel FET. Because carrier injection in a TFET relies on quantum mechanical tunneling instead of thermionic emission over a potential barrier, the TFET can theoretically achieve a SS smaller than 60 mV/dec at room temperature.

Optimization of TFET performance is still an area of active research. Many efforts have been focused on increasing the small reported ON-state drive current [36-37]. Other non-idealities such as process-induced variations and trap-assisted tunneling can significantly increase the OFF-state leakage current and degrade SS [40-43].

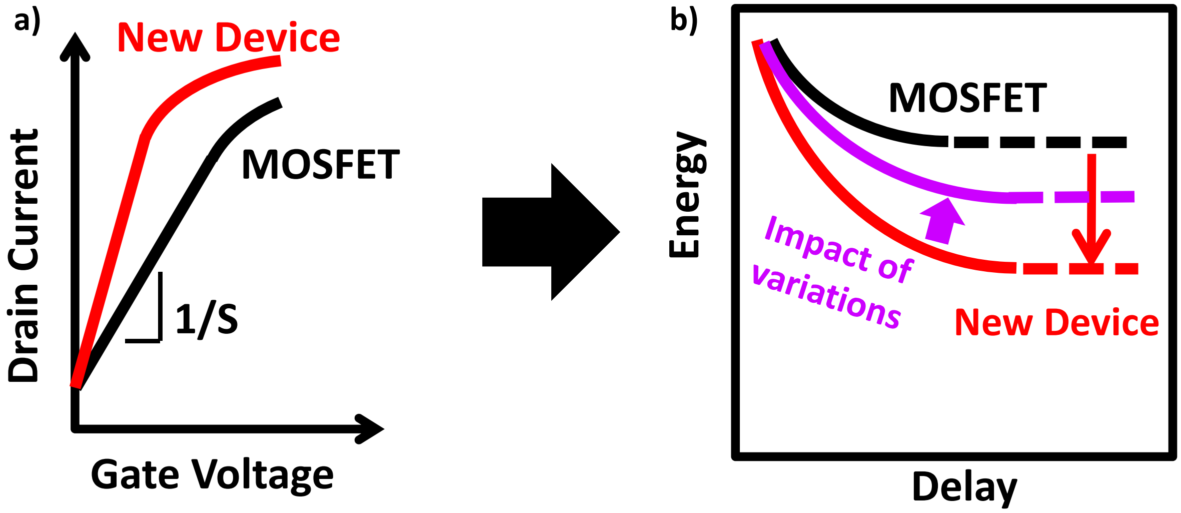


Fig. 1.12 a) A new logic device with a smaller subthreshold swing (i.e. turning on/off more steeply) as compared to the MOSFET allows for higher ON-state drive current for the same OFF-state leakage current. b) Energy per operation vs. delay can be lowered for a steep switching device. However, the impact of non-idealities such as process-induced variations will reduce the energy savings.

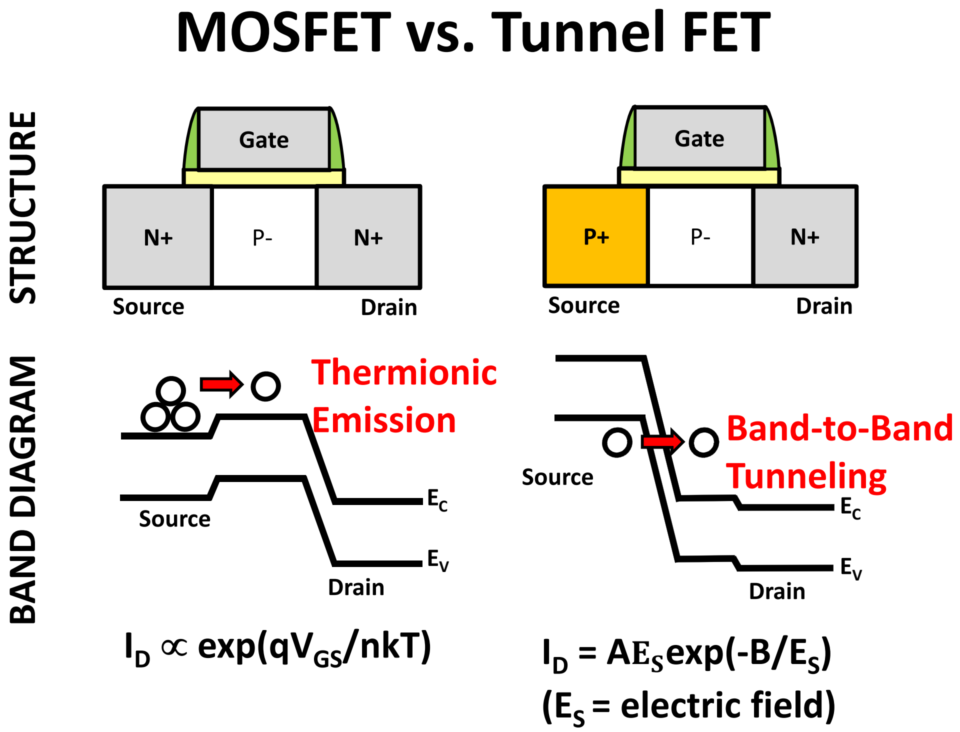


Fig. 1.13 Comparison between MOSFET and TFET structures, and their operations

1.4 Variability Sources

Process-induced variation has emerged as one of the potential limiters for Moore's Law as its effects on transistor performance has increased significantly with transistor scaling [44-45]. With the number of transistors on a chip exceeding 1 billion today, it is imperative that the actual electrical behavior of a transistor is as close as possible to the nominal characteristic as modeled for circuit design. If some transistors do not meet performance specifications, then the result can be faulty circuit operation and hence lower chip manufacturing yield. Worsening short-channel effects increase the sensitivity of transistor performance to process-induced variations and thereby compound this problem.. In today's chip designs, a large design margin or "guard band" is required in order to ensure that the circuits will still function correctly in the presence of transistor variability. Such a requirement often leads to over designing of the circuit which can result in increased power consumption or larger delay. Thus, it is imperative to understand the causes of transistor variability and how they affect transistor performance.

Variability sources are often categorized as either systematic or random [46]. Systematic variability is often dependent on the layout of the transistors and its surroundings. For example, a jog in a metal pattern can systematically lead to corner rounding after the exposure step, or a high-density pattern can have a lower etch rate resulting in etched features of different size than for a sparse pattern. Random variability, on the other hand, is more troublesome, since it can result in differences between identically drawn transistors within the same layout environment

[47-48]. The stochastic nature of random variability makes it more difficult to quantify its effects on device performance. Fig. 1.14 shows the so-called butterfly curves used to determine the read stability of six-transistor (6T) static memory (SRAM) cells [49]. The length of the side of the smaller square fitted inside one of the two lobes represents the static-noise margin (SNM), which is the largest amount of DC noise voltage that the circuit can tolerate without an error [50]. If the transistors on each side of the cell are not perfectly matched, then the two inverter voltage transfer characteristics are imbalanced, decreasing SNM. Furthermore, as the SRAM cell supply voltage V_{DD} is reduced, the SNM is reduced toward zero, indicating that there is a minimum value of V_{DD} ($V_{DD,MIN}$) for proper SRAM cell operation.

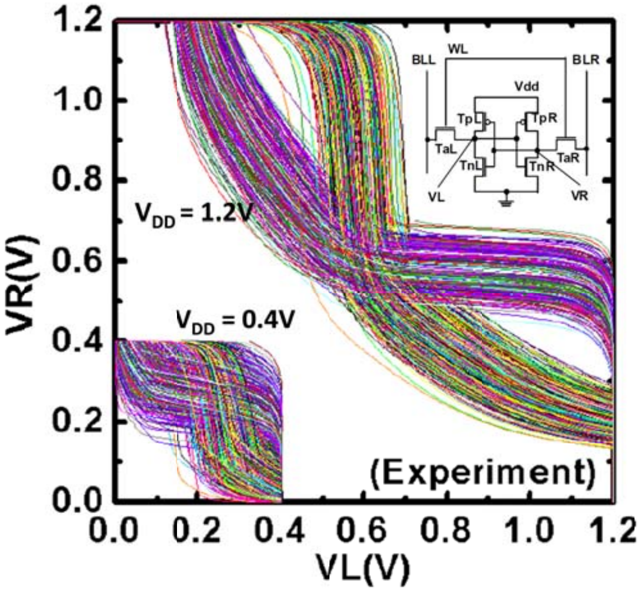


Fig. 1.14 Measured butterfly curves of 6T SRAM cells. SNM is reduced in the presence of random variability (adapted from [49]).

1.4.1 Random Dopant Fluctuation (RDF)

One of the random variability sources contributing to the variation in transistor threshold voltage V_T is random dopant fluctuation (RDF), which is caused by the variation in the number and placement of dopant atoms in the channel region of the transistor. Fig. 1.15a shows how the average number of dopants in the channel region decreases with technology scaling. The small number of dopant atoms makes V_T susceptible to even the slightest amount of dopant variation. The amount of threshold voltage variation (σV_T) can be represented analytically by Eqn.1.3:

$$\sigma V_T \propto \frac{T_{ox}}{\varepsilon_{ox}} \cdot \frac{\sqrt[4]{N_{tot}}}{\sqrt{W \cdot L}} \tag{1.3}$$

where T_{ox} = thickness of the gate oxide, ϵ_{ox} = dielectric constant of gate oxide, N_{tot} = total number of dopants, W = channel width, L = channel length [47]. Reduction in device dimensions (i.e. W and L) and increasing doping density will result in larger V_T variation. Conversely, by reducing the effective thickness of the gate oxide (e.g. by adopting a high-k dielectric), σV_T can be reduced. A more detailed analysis of V_T variability due to RDF can be performed through 3D device simulation, which randomly places dopant atoms within the transistor as shown in Fig.1.15b. Another effective method is to design a test chip with device structures to monitor random variability.

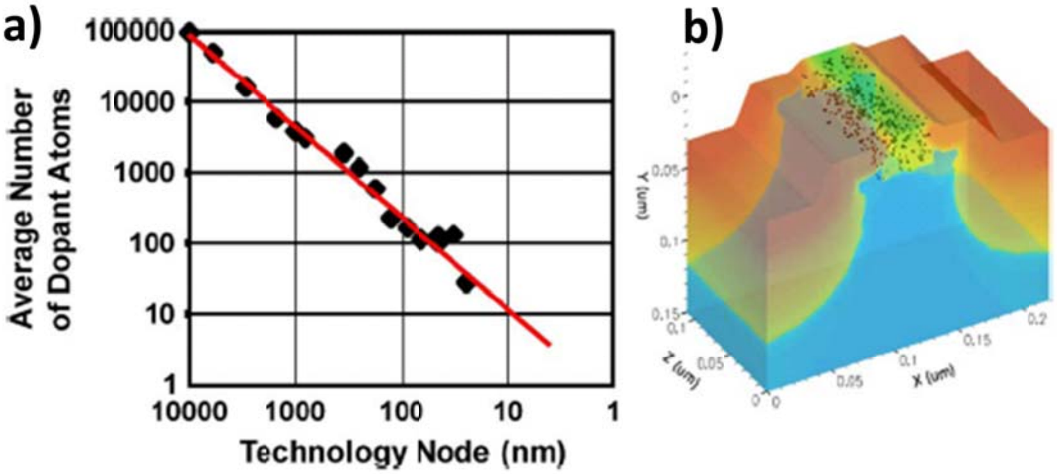


Fig. 1.15 a) Plot of average number of dopant atoms vs. technology node. b) Simulated 3D bulk-Si MOSFET structure with atomistic doping taken into account (adapted from [47]).

1.4.2 Line Edge Roughness (LER)

As transistors are scaled down, their critical dimension (CD) – the gate length – becomes so small that a slight deviation from the nominal value can have a large effect on electrical performance. One of the effects contributing to random variation in CD is line edge roughness (LER), which is caused by the granularity of the photoresist material (used to define the pattern of the gate electrodes) at the molecular level [51-52]. As the CD decreases, , LER does not decrease commensurately, which can result in large V_T variation as shown in Fig. 1.16 [44]. LER can affect both the gate length and channel width of a transistor as depicted in Fig. 1.17a. With a more advanced structure like the FinFET, fin width variation caused by LER is also becoming a major concern.

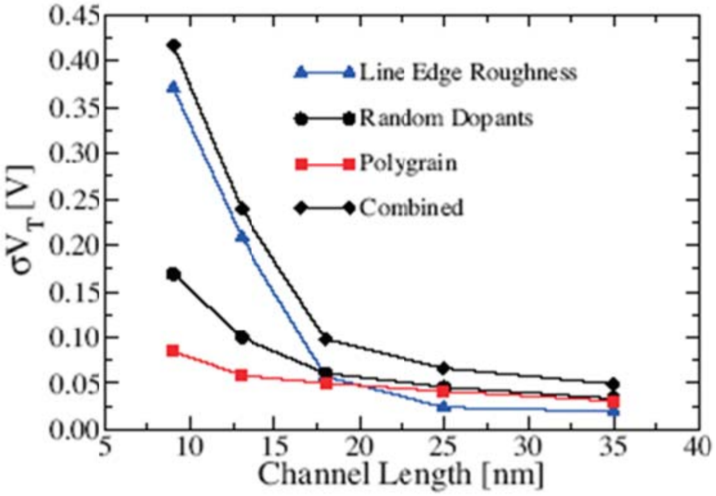


Fig. 1.16 Threshold voltage variation as a function of channel length. The contribution of gate line edge roughness is predicted to overtake RDF at channel lengths below 17 nm [44].

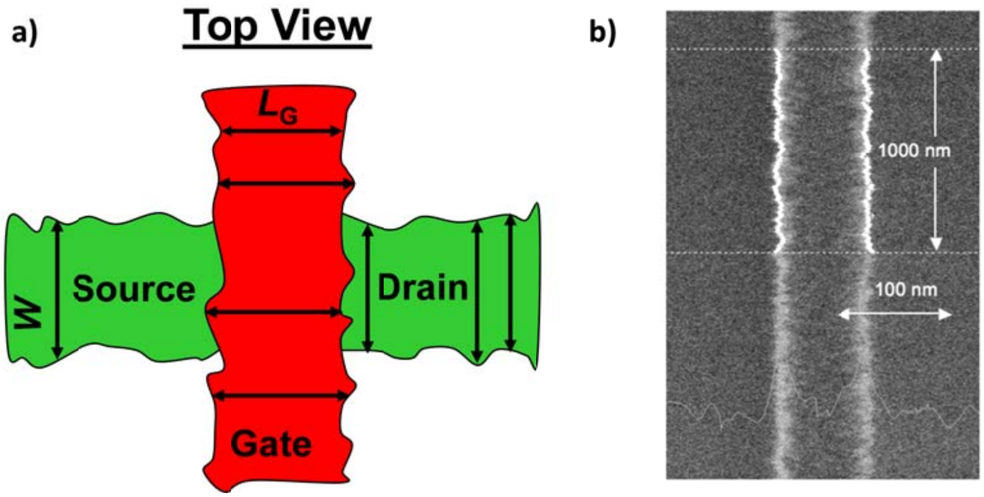


Fig. 1.17 a) Top view showing how LER can affect the gate length and channel width of a transistor. b) Scanning Electron Microscope (SEM) image of a photoresist line with LER.

Characterization of line edge roughness can be done efficiently by analyzing the top view scanning electron microscope (SEM) image of the feature as in Fig. 1.17b [53]. A more accurate measurement of LER can also be carried out using an atomic force microscope (AFM), albeit at a much slower rate [51]. There are 3 parameters that are commonly used to characterize LER: root mean square (RMS) roughness, Correlation length, and Fractal dimension [51-52, 54]. The RMS value of LER, which is often expressed in 3σ, gauges the amplitude of CD deviation. Correlation length refers to the distance between points along the feature which are correlated in the CD direction. Lastly, fractal dimension gives a degree of high-frequency fluctuation in the

LER [51]. Each of the LER parameters must meet a specification in order to enable further CD reductions as required for future technology nodes [55].

1.4.3 Gate Work Function Variation (WFV)

Another emerging source of random variability has to do with the granularity of the gate material [47]. Polysilicon or metal gate electrodes are composed of multiple crystalline grains (Fig. 1.18) [56]. Since each grain can possess a different effective work function value, there exists a gate work function variation (WFV) between transistors, leading to variation in V_T. To mitigate variability caused by WFV, researchers have proposed the use of an amorphous metal for the gate material [56]. The smaller the average grain size in the gate material, the lower the work function-induced V_T variability, as can be seen from the I_D-V_G characteristics in Fig. 1.19. Although the use of an amorphous metal gate can help suppress V_T variability, the average gate WF of the device can be shifted to an undesirable value. Therefore, it is important to simultaneously try to reduce WFV while maintaining an acceptable nominal value of V_T.

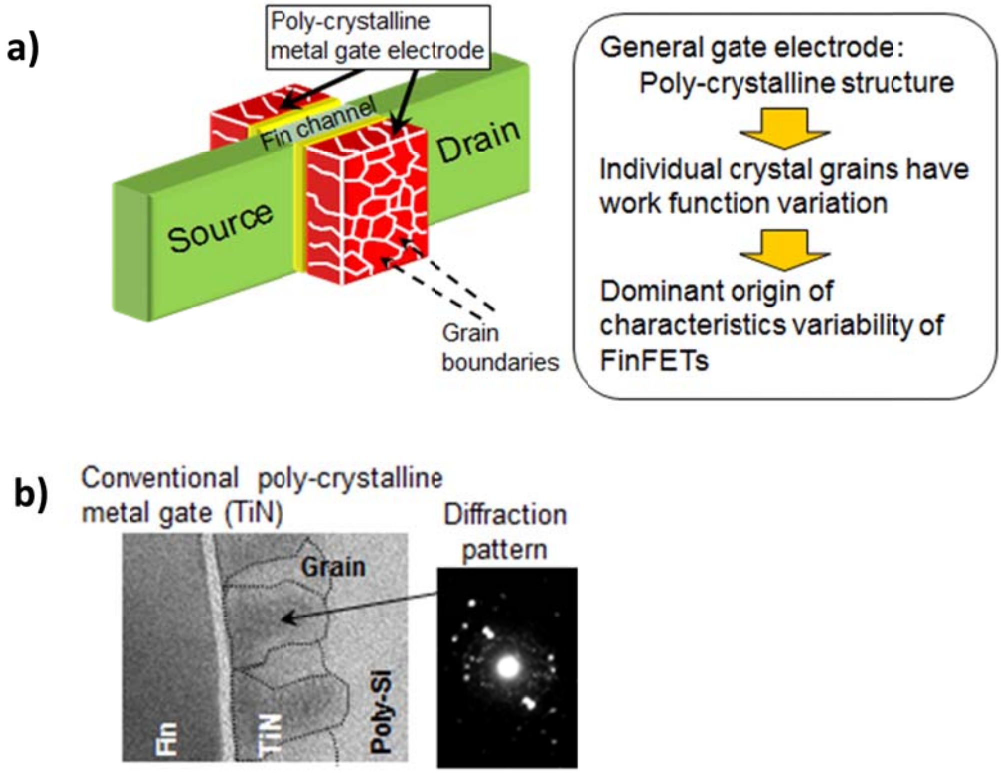

a)
Poly-crystalline metal gate electrode
Fin channel
Drain
Source
Grain boundaries
General gate electrode: Poly-crystalline structure
Individual crystal grains have work function variation
Dominant origin of characteristics variability of FinFETs
b)
Conventional poly-crystalline metal gate (TiN)
Diffraction pattern
Grain
Fin
TiN
Poly-SI

Fig. 1.18 a) Individual grains of a poly-crystalline gate metal can introduce WFV b) TEM image and a diffraction pattern of a TiN poly-crystalline gate structure showing multiple grains (adapted from [56]).

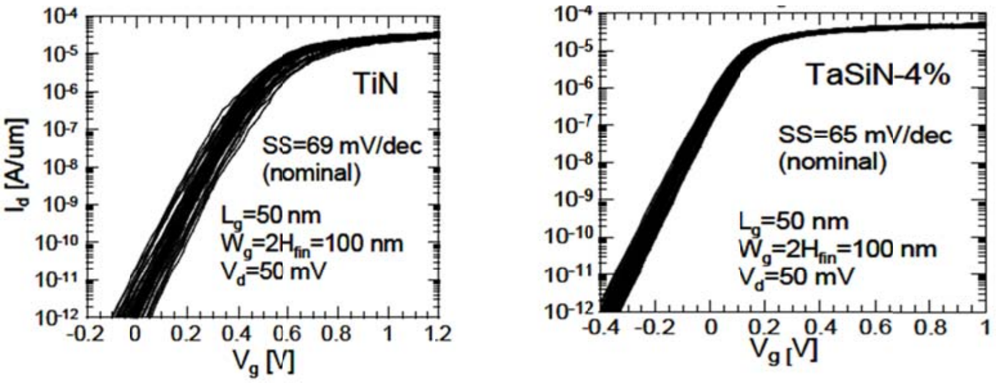


Fig. 1.19 Measured I_D – V_G curves of transistors with poly crystalline TiN (Left) vs. amorphous TaSiN (Right) gate material. The introduction of TiSiN helps reduce σV_T due to WFV, but it also lowers the average V_T by a significant amount [56].

1.5 Research Objectives and Thesis Overview

In **Chapter 2**, the use of super-steep retrograde (SSR) channel doping to extend planar bulk-Si MOSFET scalability is investigated. Three-dimensional device simulations are performed to compare the performance of MOSFETs with SSR vs. uniform channel doping, for physical gate length L_g = 28nm. Random dopant fluctuation is introduced to examine the effectiveness of SSR doping in mitigating RDF-induced variability. The implication for 6T SRAM cell yield is also analyzed and the result indicates that SSR channel doping has the potential to lower $V_{DD,MIN}$ of SRAM cell, translating to a significant power savings.

In **Chapter 3**, a physically-based variability model is developed to explain variations in threshold voltage (V_T) and drain induced barrier lowering (DIBL), and their correlations for SRAM and analog transistors fabricated in a 32nm HKMG technology. Inputs to the model rely on measurements of forward (F) and reverse (R) mode characteristics of transistors in a mismatch pair. Modeling results are validated using SRAM and analog devices. Asymmetric and symmetric variation components of V_T and DIBL variability are extracted using the model. Asymmetric variation is found to be a major component responsible for the higher σV_T mismatch in the saturation region of transistor operation as compared to the linear region of operation, and higher DIBL variability.

In **Chapter 4**, the design of a test chip to study the impact of device variability in 28nm planar bulk and FD-SOI MOSFETs is discussed. A device characterization array including transistors in mismatch pairs and different layout proximity is used to study the impact of random and systematic variability, respectively. Padded-out cells in an SRAM array are designed for a 16nm FD-SOI CMOS technology to study the impact of variability. The padded-out cell design allows for direct correlation between SRAM performance metrics and transistor characteristics. A built-in stress mode is also included in the padded-out SRAM design to allow for characterization of SRAM cell performance degradation due to NBTI/PBTI and RTN.

In **Chapter 5**, a comprehensive study of process-induced variability due to RDF and LER is investigated for planar and raised Germanium-Source Tunnel FETs. Device characteristics are studied via three-dimensional device simulation calibrated to experimental data. The contributions to RDF-induced V_T variation due to atomistic doping in the different regions of the device are identified. Different source-edge roughness profiles of a TFET are considered to assess the impact of LER. The combined effect of RDF- and LER- induced variations on V_T, I_{ON}, and I_{OFF} is analyzed. The energy *vs.* delay performance of a TFET accounting for variability is benchmarked against that of a MOSFET.

In **Chapter 6**, the contributions of this dissertation are summarized and suggestions for future work are offered.

1.6 References

[1] N. H. E. Weste, D. M. Harris, CMOS VLSI Design: A circuits and systems perspective, 4th ed., Addison-Wesley, 2011.

[2] J. M. Rabaey, A. Chandrakasan, B. Nikolic, Digital Integrated Circuit: A Design Perspective, 2nd ed., Pearson Education, 2003.

[3] G. E. Moore, "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff.," *Solid-State Circuits Society Newsletter, IEEE* , vol.11, no.5, pp.33,35, Sept. 2006.

[4] S. E. Thompson, S. Parthasarathy, "Moore's law: the future of Si microelectronics," *Materials Today*, Volume 9, Issue 6, June 2006, Pages 20-25

[5] R.R. Schaller, "Moore's law: past, present and future," *Spectrum, IEEE* , vol.34, no.6, pp.52,59, Jun 1997

[6] R.H. Dennard ; F.H. Gaensslen; H.-N. Yu; V.L. Rideout; E. Bassous; A. R. Leblanc, "Design Of Ion-implanted MOSFET's with Very Small Physical Dimensions," *JSSC*, vol. SC-9, no. 5, Oct. 1974, pp 256-268.

[7] W. Arden, M. Brillouet, P. Cogez, M. Graef, B. Huizing, R. Mahnkopf, "More-than-Moore White Paper," *ITRS* 2010. Available: <http://www.itrs.net/Links/2010ITRS/IRC-ITRS-MtM-v2%203.pdf>

[8] Y. Taur; D. A. Buchanan; W. Chen; D.J. Frank; K.E. Ismail; L. Shih-Hsien; G.A, Sai-Halasz; R.G. Viswanathan ; H.-J.C. Wann; S.J. Wind; Hon-Sum Wong, "CMOS scaling into the nanometer regime," *Proceedings of the IEEE* , vol.85, no.4, pp.486,504, Apr 1997.

[9] P. Packan, “Device and Circuit Interactions,” *IEEE International Electron Device meeting (IEDM ’08)* Short Course: Performance Boosters for Advanced CMOS Devices, 2007.

[10] B. Meyerson, *Semico Impact Conference*, Taiwan, 2004.

[11] S. Yang, L. G. Lin, M. Han, D. Yang, J. Wang, K. Mahmood, T. Song, D. Yuan, D. Seo, M. Pedrali-Noy, D. Alladi, S. Wadhwa, X. Bai, L. Dai, S. S. Yoon, E. Terzioglu, S. Bazarjani, G. Yeap, “High Performance Mobile SoC Design and Technology Co-Optimization to Mitigate High-K Metal Gate Process Induced Variations” in *VLSI Symp. Tech.* Dig., 2014.

[12] V. Moroz, “Transition From Planar MOSFETs to FinFETs and its Impact on Design and Variability”, *Berkeley Seminar*, Oct. 2011.

[13] S.E. Thompson; G. Sun; Y.S. Choi; T. Nishida, “Uniaxial-process-induced strained-Si: extending the CMOS roadmap,” *Electron Devices, IEEE Transactions on* , vol.53, no.5, pp.1010,1020, May 2006.

[14] S. Ito, H. Namba, K. Yamaguchi, T. Hirata, K. Ando, S. Koyama, S. Kuroki, N. Ikezawa, T. Saitoh, T. Horiuchi, “Mechanical Stress Effect of Etch-Stop Nitride and its Impact on Deep Submicron Transistor Design,” *IEDM Tech. Dig.*, 2000.

[15] K. Ota, K. Sugihara, H. Sayama, T. Uchida, H. Oda, T. Eimori, H. Morimoto, Y. Inoue, “Novel Locally Strained Channel Technique for High Performance 55nm CMOS,” *IEDM Tech. Dig.,* 2002.

[16] T. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, K. Johnson, C. Kenyon, J. Klaus, B. McIntyre, K. Mistry, A. Murthy, J. Sandford, M. Silberstein, S. Sivakumar, P. Smith, K. Zawadzki, S. Thompson, M. Bohr, “A 90nm High Volume Manufacturing Logic Technology Featuring Novel 45nm Gate Length Strained Silicon CMOS Transistors,” *IEDM Tech. Dig.*, 2003.

[17] C. Auth, A. Cappellani, J.-S. Chun, A. Dalis, A. Davis, T. Ghani, G. Glass, T. Glassman, M. Harper, M. Hattendorf, P. Hentges, S. Jaloviar, S. Joshi, J. Klaus, K. Kuhn, D. Lavric, M. Lu, H. Mariappan, K. Mistry, B. Norris, N. Rahhal-orabi, P. Ranade, J. Sandford, L. Shifren, V. Souw, K. Tone, F. Tambwe, A. Thompson, D. Towner,T. Troeger, P. Vandervoorn, C. Wallace, J. Wiedemer, C. Wiegand, “45nm High-k+Metal Gate Strain-Enhanced Transistors,” *Symp. on VLSI Tech Dig.*, 2008.

[18] S.E. Thompson, P.A. Packan, M.T. Bohr, “Linear versus saturated drive current: Tradeoffs in super steep retrograde well engineering,” in *VLSI Symp. Tech. Dig.*, 1996, pp. 12 -13.

[19] A. Hokazono ; H. Itokawa; N. Kusunoki; I. Mizushima; S. Inaba; S. Kawanaka; Y. Toyoshima, “Steep channel & Halo profiles utilizing boron-diffusion-barrier layers (Si:C) for 32 nm node and beyond,” in *VLSI Symp. Tech. Dig.*, 2008, pp. 112–113.

[20] C. Auth; C. Allen; A. Blattner; D. Bergstrom; M. Brazier; M. Bost; M. Buehler; V. Chikarmane; T. Ghani; T. Glassman; R. Grover; W. Han; D. Hanken; M. Hattendorf; P. Hentges; R. Heussner; J. Hicks; D. Ingerly; P. Jain; S. Jaloviar; R. James; D. Jones; J. Jopling; S. Joshi; C. Kenyon; H. Liu; R. McFadden; B. Mcintyre; J. Neirynck; C. Parker; L. Pipes; I. Post; S. Pradhan; M. Prince; S. Ramey; T. Reynolds; J. Roesler; J. Sandford; J. Seiple; P. Smith; C. Thomas; D. Towner; T. Troeger; C. Weber; P. Yashar; K. Zawadzki; K. Mistry, “A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors,” *VLSI Technology (VLSIT), 2012 Symposium on* , vol., no., pp.131,132, 12-14 June 2012.

[21] W. Shien-Yang; C,Y, Lin; M.C. Chiang; J.J. Liaw; J.Y. Cheng; S.H. Yang; M. Liang; T. Miyashita; C.H. Tsai; B.C. Hsu; H.Y. Chen; T. Yamamoto; S.Y. Chang; V.S. Chang; C.H. Chang; J.H. Chen; H.F. Chen; K.C. Ting; Y.K. Wu; K.H. Pan; R.F. Tsui; C.H. Yao; P.R. Chang; H.M. Lien; T.L. Lee; H.M. Lee; W. Chang; T. Chang; R. Chen; M. Yeh; C.C. Chen; Y.H. Chiu; Y.H. Chen; H.C. Huang; Y.C. Lu; C.W. Chang; M.H. Tsai; C.C. Liu; K.S. Chen; C.C. Kuo; H.T. Lin; S.M. Jang; Y. Ku, “A 16nm FinFET CMOS technology for mobile SoC and computing applications,” *Electron Devices Meeting (IEDM), 2013 IEEE International* , vol., no., pp.9.1.1,9.1.4, 9-11 Dec. 2013

[22] ST Microelectronics. <http://www.st.com/web/en/about_st/fd-soi.html>

[23] L. Le Pailleur, “28nm FD-SOI Industrial Solution: Overview of Silicon Proven Key Benefits” *FDSOI- Workshop* 2013 Kyoto, Japan.

[24] C. Fenouillet-Beranger; P. Perreau; T. Benoist; C. Richier; S. Haendler; J. Pradelle; J. Bustos; P. Brun; L. Tosti; O. Weber; F. Andrieu; B. Orlando; D. Pellissier-Tanon; F. Abbate; C. Pvichard; R. Beneyton; M. Gregoire; J. Ducote; P. Gouraud; A. Margain; C. Borowiak; R. Bianchini; N. Planes; E. Gourvest; K.K. Bourdelle; B.Y. Nguyen; T. Poiroux; T. Skotnicki; O. Faynot; F. Boeuf, “Impact of local back biasing on performance in hybrid FDSOI/bulk high-k/metal gate low power (LP) technology,” *Ultimate Integration on Silicon (ULIS), 2012 13th International Conference on* , vol., no., pp.165,168, 6-7 March 2012.

[25] N. Planes; O. Weber; V. Barral; S. Haendler; D. Noblet; D. Croain; M. Bocat; P. Sassoulas; X. Federspiel; A. Cros; A. Bajolet; E. Richard; B. Dumont; P. Perreau;

D. Petit; D. Golanski; C. Fenouillet-Beranger; N. Guillot; M. Rafik; V. Huard; S. Puget; X. Montagner; M.-A. Jaud; O. Rozeau; O. Saxod; F. Wacquant; F. Monsieur; D. Barge; L. Pinzelli; M. Mellier; F. Boeuf; F. Arnaud; M. Haond, "28nm FDSOI Technology Platform for High-Speed Low-Voltage Digital Applications," in *VLSI Symp. Tech. Dig.*, 2012.

[26] C. Hu, Modern Semiconductor Devices for ICs, Pearson, 2010.

[27] X. Huang; W.-C. Lee; C. Kuo; D. Hisamoto; L. Chang; J. Kedzierski; E. Anderson; H. Takeuchi; Y.-K. Choi; K. Asano; V. Subramanian; T.-J. King; J. Bokor; C. Hu, "Sub 50-nm FinFET: PMOS," *Electron Devices Meeting, 1999. IEDM '99. Technical Digest. International*, vol., no., pp.67,70, 5-8 Dec. 1999.

[28] D. Hisamoto; W.-C. Lee; J. Kedzierski; H. Takeuchi; K. Asano; C. Kuo; E. Anderson; T.-J. King; J. Bokor; C. Hu, "FinFET-a self-aligned double-gate MOSFET scalable to 20 nm," *Electron Devices, IEEE Transactions on* , vol.47, no.12, pp.2320,2325, Dec 2000.

[29] J.-W. Yang; J.G. Fossum, "On the feasibility of nanoscale triple-gate CMOS transistors," *Electron Devices, IEEE Transactions on* , vol.52, no.6, pp.1159,1164, June 2005.

[30] M. Bohr, K. Mistry, "Intel's Revolutionary 22 nm Transistor Technology," *Intel Presentation*, May 2011.

[31] D. Fried, T. Hoffmann, B.-Y. Nguyen, S. Samavedam, "Comparison study of FinFETs: SOI vs. Bulk," *SOI Industry Consortium*, Oct 2009.

[32] T.-J. King, "FinFET History, Fundamentals and Future," *VLSI Symp. Technology* Short Course 2007.

[33] S.H. Rasouli; H.F. Dadgour; K. Endo; H. Koike; K. Banerjee, "Design Optimization of FinFET Domino Logic Considering the Width Quantization Property," *Electron Devices, IEEE Transactions on* , vol.57, no.11, pp.2934,2943, Nov. 2010.

[34] T.-J. King, N. Xu, "FinFET versus UTVV SOI for RF/Analog Applications," ISSCC Forum F6: Mixed-Signal/RF Design and Modeling in Next-Generation CMOS. Feb. 2013.

[35] A. M. Ionescu and H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches" *Nature* 479(7373): 329-337, Nov 2011.

[36] G. Dewey, et al., "Fabrication, Characterization, and Physics of III-V Heterojunction Tunneling field Effect Transistors (H-TFET) for Steep Sub-Threshold Swing," in *IEDM Tech. Dig.,* pp. 33.6.1-33.6.4, Dec. 2011.

[37] A. Villalon, et al., "Strained Tunnel FETs with record I_{ON}: First demonstration of ETSOI TFETs with SiGe channel and RSD," in *VLSI Symp. Tech. Dig.,* June 2012, pp. 49-50.

[38] S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices (Issue Cover Story)," *Nanoletters*, vol. 8, no. 2, pp. 405-410, Feb. 2008.

[39] R. Nathanael, V. Pott, H. Kam, J. Jeon and T. -J. K. Liu, "4-terminal relay technology for complementary logic", in *IEDM Tech. Dig*., pp. 223-226, 2009.

[40] N. Damrongplasit, C. Shin, S. H. Kim, R. Vega, and T. –J. K. Liu, "Study of Random Dopant Fluctuation Effects in Germanium-Source Tunnel FETs," *IEEE Trans. Electron Devices*, vol. 58, no. 10, Oct. 2011, pp. 3541-3548.

[41] N. Damrongplasit, S. H. Kim, C. Shin, and T. –J. K. Liu, ", Impact of Gate Line-Edge Roughness (LER) Versus Random Dopant Fluctuations (RDF) on Germanium-Source Tunnel FET Performance," *IEEE Trans. Nanotechnology*, vol. 12, issue. 6, Nov. 2013, pp. 1061-1067.

[42] F. Conzatti, M. G. Pala, and D. Esseni, "Surface-Roughness-Induced Varaibility in Nanowire InAs Tunnel FETs," *IEEE Electron Device Letters,* vol.33, no. 6, Jun. 2012, pp. 806-808.

[43] A.M. Walke; A.S. Verhulst; A. Vandooren; D. Verreck; E. Simoen; V.R. Rao; G. Groeseneken; N. Collaert; A.V.Y. Thean, "Part I: Impact of Field-Induced Quantum Confinement on the Subthreshold Swing Behavior of Line TFETs," *Electron Devices, IEEE Transactions on* , vol.60, no.12, pp.4057,4064, Dec. 2013.

[44] A. Asenov, "Simulation of statistical variability in nano MOSFETs," *IEEE Symp. VLSI Technol., Dig. Tech. Papers*, Jun. 2007, pp. 86–87.

[45] A. Asenov, A. Brown, J. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuation in decananometer and nanometer-scale MOSFETs", *IEEE Trans. on Electron Devices*, vol. 50, 2003, pp.1837-1852.

[46] K. J. Kuhn, "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS", *IEDM Tech. Dig*., pp.471 -474 2007.

[47] K. J. Kuhn, M. D. Giles, D. Becher, P. Koler, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai, “Process technology variation,” IEEE Trans. Electron Devices, vol. 58, no. 8, pp. 2197–2208, Aug. 2011

[48] K. Takeuchi, A. Nishida, T. Hiramoto, “Random Fluctuations in Scaled MOS Devices,” *Simulation of Semiconductor Processes and Devices, 2009. SISPAD '09. International Conference on* , vol., no., pp.1,7, 9-11 Sept. 2009

[49] T. Hiramoto, “Measurements and Characterization of Statistical Variability,” *SISPAD* workshop 2010.

[50] E. Seevinck; F.J. List; J. Lohstroh, “Static-noise margin analysis of MOS SRAM cells,” *Solid-State Circuits, IEEE Journal of* , vol.22, no.5, pp.748,754, Oct 1987.

[51] V. Constantoudis, G.P. Patsis, L.H.A. Leunissen, E. Gogolides, “Line edge roughness an critical dimension variation: Fractal characterization and comparison using model functions,” *J. Vac. Sci. Technol*. B 22, 1974 (2004)

[52] Y. Ma; H. J. Levinson; T. Wallow, “Line edge roughness impact on critical dimension variation,” *Proc. SPIE 6518, Metrology, Inspection, and Process Control for Microlithography* XXI, 651824 (April 05, 2007)

[53] A. Hiraiwa, A. Nishida, “Statistical- and image-noise effects on experimental spectrum of line-edge and line-width roughness,” *J. Micro/Nanolith. MEMS MOEMS*, Dec. 2010.

[54] A. Asenov, S. Kaya, and A. R. Brown, “Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness,” *IEEE Trans. Electron Devices*, vol 50, no. 5, pp. 1254-1260, May 2003.

[55] International Technology Roadmap for Semiconductor (ITRS) report, 2013.

[56] T. Matsukawa; Y. Liu; W. Mizubayashi; J. Tsukada; H. Yamauchi; K. Endo; Y. Ishikawa; S. O'uchi; H. Ota; S. Migita; Y. Morita; M. Masahara, “Suppressing Vt and Gm variability of FinFETs using amorphous metal gates for 14 nm and beyond,” *Electron Devices Meeting (IEDM), 2012 IEEE International* , vol., no., pp.8.2.1,8.2.4, 10-13 Dec. 2012.

Chapter 2

Comparative Study of Uniform vs Super-Steep Retrograde Channel MOSFET and Implications for 6T SRAM Yield

2.1 Introduction

Short-channel effects and variability in planar bulk silicon (bulk-Si) MOSFET performance worsen as the gate length (L_g) is scaled down below 30 nm [1]. For the conventional six-transistor (6-T) SRAM cell design, variability-induced transistor mismatch results in cell imbalance which limits reductions in cell operating voltage [2]. These challenges can be overcome by adopting thin-body MOSFET structures, such as the FinFET [3] or the fully depleted silicon-on-insulator (FD-SOI) MOSFET [4], which can have superior electrostatic integrity. However, such structures require greater process complexity and/or more expensive starting substrates, so that increased manufacturing cost is a concern. Since price is a key factor for mobile electronics applications, planar bulk-Si CMOS technology still can be competitive because of its lower process and/or substrate cost.

Super-steep retrograde (SSR) channel doping has been proposed to extend the scalability of planar bulk-Si CMOS technology [5], and has been investigated in MOSFETs with relatively long channel lengths by today's standards [6-9]. Conventional doping approaches mostly utilize ion implantation, which unavoidably results in dopants near to the channel-region surface, since the dopant atoms are introduced through the surface. Recently developed epitaxial silicon technology can alleviate this effect, since undoped silicon can be grown on diffusion barrier layers, such as carbon-doped silicon [10-11] or oxygen partial monolayers [12-13], after introduction of the ground-plane dopant atoms. In particular, the use of oxygen partial monolayers [12] has not only the dopant diffusion blocking effect but also a dopant pile-up effect below the inserted layers and thus shows promise for achieving the ideal SSR profile. In this letter, the benefit of SSR channel doping over uniform channel doping (UD) for reducing variability in planar bulk-Si MOSFET performance and improving SRAM yield is studied via TCAD simulation of three-dimensional (3-D) devices with 28 nm gate length (L_g).

2.2 Device Simulation Approach

Device simulations are performed with Sentaurus using drift-diffusion transport, density gradient model for quantum confinement, bandgap narrowing effect, Philips and high-field degradation models for mobility [14]. The N-channel (NMOS) and P-channel (PMOS) transistor designs (gate offset spacer width (L_{OS}), source/drain (S/D) extension junction depth ($X_{j,ext}$), S/D offset spacer width (L_{sd}), and S/D-extension peak active dopant concentration (N_{ext})) are optimized for maximum on-state drive current ($I_{D,Sat}$) with off-state leakage (I_{OFF}) $\leq$ 1 nA/um for a supply voltage (V_{DD}) of 1V. Performance boosters such as embedded S/D stressors and contact etch stop liners (CESL) are not included in the device simulations. Fig. 2.1(a) shows a simulated MOSFET cross-section.

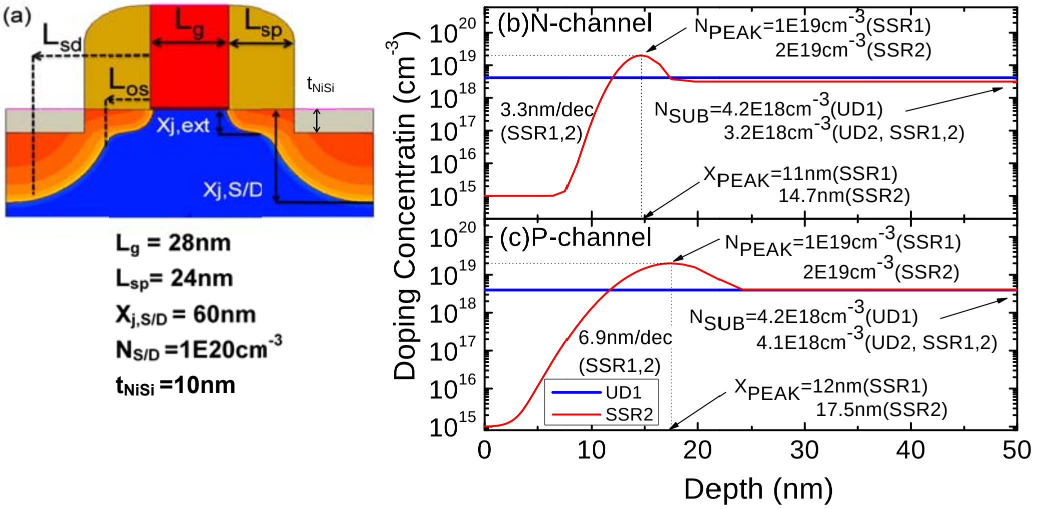


Fig 2.1 a) Simulated cross-section of a MOSFET with UD profile. Channel doping profiles used in this work for b) n-channel and c) p-channel MOSFETs. The S/D extension doping profiles were optimized to achieve maximum $I_{D,Sat}$ with $I_{OFF} \leq$ 1 nA/um: $X_{j,ext}$ = 16nm(N-UD1), 14nm(N-UD2), 11.4nm(N-SSR1), 14.7nm(N-SSR2); 16nm(P-UD1), 16nm(P-UD2), 12.8nm(P-SSR1), 18nm(P-SSR2). N_{ext} = 1E20cm^{-3}(N-UD1), 2.7E19cm^{-3}(N-UD2, N-SSR1, N-SSR2); 1E20cm^{-3}(P-UD1), 1E19cm^{-3}(P-UD2,P-SSR1,P-SSR2).

The equivalent gate-oxide thickness (EOT) is 1.6 nm, and the gate material is polycrystalline-silicon doped n-type for NMOS and p-type for PMOS with an active dopant concentration of 2.5×10^{20} cm^{-3}. The thickness of the nickel-silicide source/drain contact layer is 10 nm, and the contact resistivity is 10^{-8} ohm-cm^2.

Figs. 2.1b and 2.1c show the channel doping profiles used for NMOS and PMOS devices in this work, respectively. Two levels of UD (*i.e.* UD1 and UD2) are considered in this work. For UD1, the device design is optimized under only the constraint $I_{OFF} \leq$ 1 nA/um. For UD2, the device design is optimized under the constraints $I_{OFF} \leq$ 1 nA/um and drain-induced barrier lowering (DIBL) $\leq$ 0.1 V/V. It should be noted that the uniform doping profile approximates the case of a device formed with halo implants, since the halo doping profiles can overlap underneath the channel surface. Furthermore, after thermal annealing, the channel doping profile becomes even more uniform. For example, process simulation shows that the gradient of the halo doping profile required to match the peak depth of the SSR channel in this study is 238 nm/dec at the channel center after an 850^{0}C 30s annealing step as shown in Fig. 2.2.

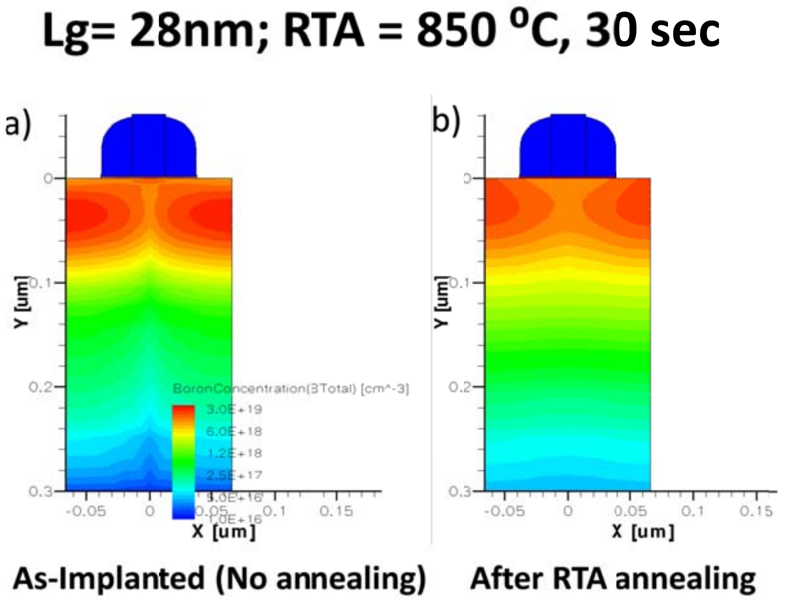


Fig. 2.2 Process simulation showing Boron doping profile a) as-implanted (no annealing) and b) after rapid thermal annealing at 850 C for t = 30 s.

Device optimization is also performed for MOSFETs with SSR channel doping profiles. The peak location and peak level of the SSR profile (a Gaussian distribution which decays toward the channel surface down to 10^{15} cm^{-3}, and decays toward the Si substrate until it merges with the background UD2 concentration) are co-optimized to achieve maximum I_{ON} under the constraints $I_{OFF} \leq 1$ nA/um and DIBL ≤ 0.1 V/V. The SSR gradient is assumed to be 3.3 nm/dec and 6.9 nm/dec in the NMOS and PMOS devices, respectively. Such steep channel doping profiles can be obtained through the use of a Si:C diffusion blocking layer [10-11], or by selective epitaxy of silicon incorporating partial monolayers of oxygen atoms after the standard shallow trench isolation (STI) and well doping processes; the oxygen in non-substitutional sites serves to impede dopant diffusion [12]. Fig.2.3 shows the effect of oxygen insertion layer on energy potential of a silicon lattice. The transmission electron microscopy (TEM) image of a transistor with oxygen-insertion layer, and the corresponding secondary ion mass spectrometry (SIMS) profile of the channel doping are shown in Fig. 2.4. Moreover, the use of oxygen partial monolayers has been shown to be beneficial for enhancing mobility and improving electrostatic integrity [13]. Two SSR designs are considered in this work. For SSR1, the channel depletion region extends below the SSR peak. For SSR2, the depletion region terminates at the depth of the SSR peak.

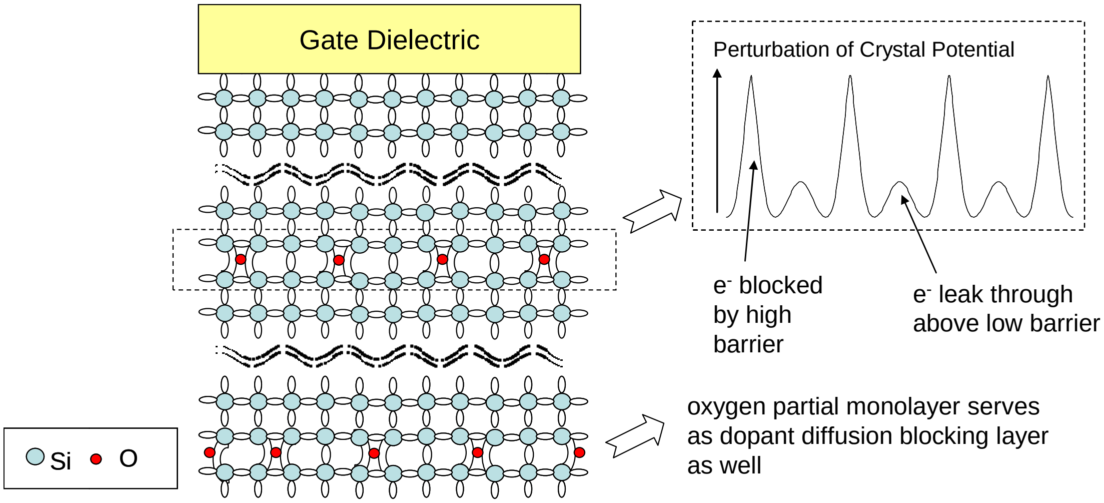

Gate Dielectric
Perturbation of Crystal Potential
e- blocked by high barrier
e- leak through above low barrier
oxygen partial monolayer serves as dopant diffusion blocking layer as well
Si
O

Fig. 2.3 Introduction of partial monolayer of oxygen results in the perturbation of silicon crystal potential.

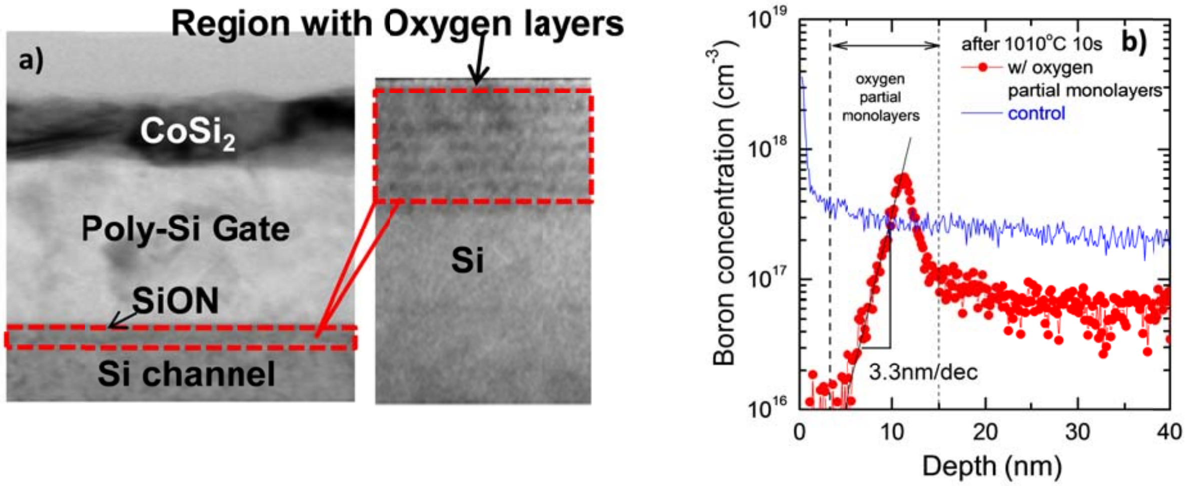


Fig. 2.4 a) TEM cross-sectional view of a bulk-Si MOSFET with partial monolayers of oxygen inserted b) SIMS profile of Boron concentration showing a 3.3 nm/dec super steep doping profile.

Variation due to random dopant fluctuations (RDF) is taken into account by simulating 200 devices with atomistic doping profiles [15], for the NMOS and PMOS transistors of each design. The minimum operating voltage ($V_{DD,min}$) for a logic circuit comprising 100M logic stages is estimated following the methodology in [16].

2.3 Device Simulation Result

Table 1 summarizes the performance parameters for the different transistor designs. UD devices generally show higher $I_{D,Sat}$ as compared to SSR devices, due to smaller electrical channel length (L_{EFF}) in the UD devices and larger body factor for the SSR devices [6]. DIBL is much lower for SSR devices, in particular SSR2. SSR2 shows the worst subthreshold swing (SS), however, due to its larger depletion capacitance (C_{dep}).

TABLE 2.1
COMPARISON OF TRANSISTOR PERFORMANCE PARAMETERS

	NMOS UD 1	PMOS UD 1	NMOS UD 2	PMOS UD 2	NMOS SSR 1	PMOS SSR 1	NMOS SSR 2	PMOS SSR 2
L_{EFF} †(nm)	34.5	34.9	39.4	71.8	48.2	67.2	48.2	66.9
$I_{D,Sat}$ (A/um)	4.76E-04	2.52E-04	4.37E-04	1.43E-04	4.28E-04	2.27E-04	3.84E-04	1.96E-04
I_{OFF} (A/um)	1.08E-09	5.79E-10	8.21E-10	1.33E-11	8.74E-10	8.80E-10	9.15E-10	8.46E-10
V_T Sat (mV)	335	-380	335	-499	321	-347	343	-370
V_T Lin(mV)	467	-536	435	-597	393	-446	389	-429
SS (mV/dec)	91	94	88	86	86	89	93	95
DIBL(mV/V)	139	164	105	103	76	104	48	62
$\sigma V_{T,}$ Sat (mV)	49.6	52.9	43.9	45.6	32.8	40.0	25.4	22.4
Body Factor ($V^{1/2}$)	0.55	0.54	0.48	0.55	0.85	0.71	1.20	1.20
Logic $V_{DD,MIN}$ (mV)	380		364		296		225	

† L_{EFF} is defined as the lateral distance at which S/D doping decays to 2 x 10^{19} cm^{-3}

V_T roll-off curves are shown in Fig 2.5 (L_g is varied from 12 nm to 500 nm; the doping profiles are unchanged.) The poor V_T roll-off characteristics for the uniformly doped channels are attributed to relatively thick EOT value used in this work. The short-channel effect is best suppressed with the SSR2 channel doping profile for the same EOT, indicating that SSR channel doping can extend poly-Si/SiON gate stack technologies beyond the 28nm node.

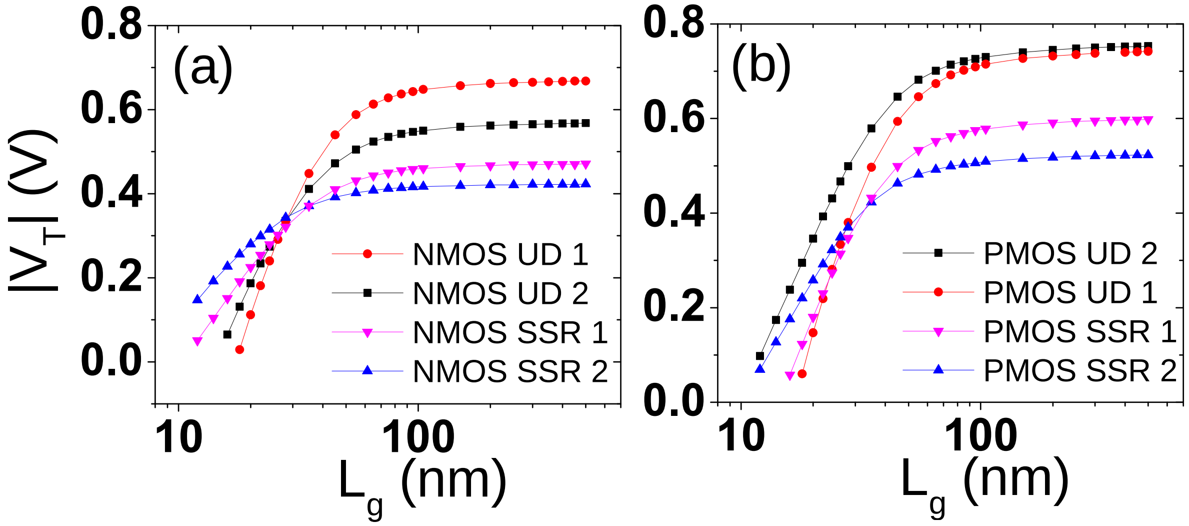


Fig. 2.5 V_T *vs.* L_g plots for a) NMOS and b) PMOS devices. V_T is defined at a constant current (I_o = 100 nA·W/L, V_{DS} = 1V).

$I_{D,Sat}$ *vs.* I_{OFF} scatter plots are shown in Fig. 2.6. The distribution of performance for UD1 devices shows that they generally achieve higher $I_{D,Sat}$ as compared to UD2 devices with the same I_{OFF}. This is due to larger parasitic series resistance and longer L_{EFF} in the UD2 devices. Although SSR devices show less variation due to better suppression of SCE, they also have degraded $I_{D,Sat}$ as compared with UD devices: at I_{OFF} = 1 nA/um, the average value of $I_{D,Sat}$ is lower for SSR2 than for UD1, by 7% for NMOS and 20% for PMOS.

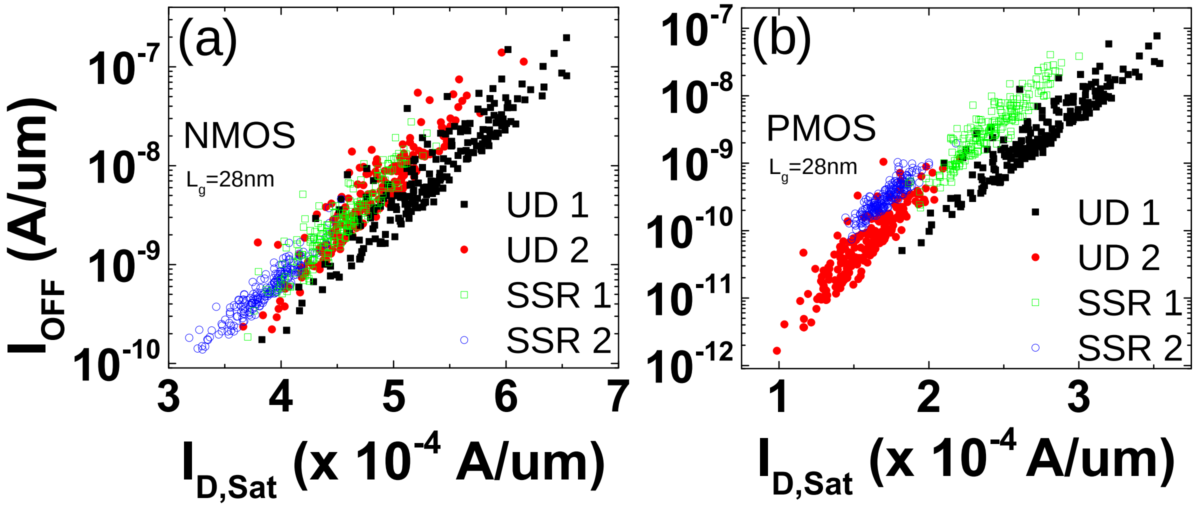


Fig. 2.6 $I_{D,Sat}$-I_{OFF} scatter plots due to RDF for a) NMOS and b) PMOS devices. Nominal device width=50 nm.

A comparison of V_T variation between SSR1 and SSR2 devices shows that σV_T is most effectively suppressed when the channel depletion region does not extend beyond the SSR peak.

However, the body factor is larger and hence $I_{D,sat}$ is smaller in this case (SSR2).

Factors which determine logic $V_{DD,min}$ are SS, DIBL, βp/βn (transistor drive current ratio between PMOS and NMOS) balance, and V_T variation. SSR devices are projected to have lower logic $V_{DD,min}$ due to smaller V_T variation and lower DIBL, as summarized in Table 2.1.

2.4 6T SRAM Yield Analysis

Read static noise margin (SNM) [17] and writeability current ($I_{W,}$ extracted from the write N-curve) [18] are used to gauge the read stability and write stability of an SRAM cell. To estimate 6-T SRAM cell performance and yield, physically-based analytical models [2] are calibrated to current *vs.* voltage characteristics obtained from 3-D device simulations, for both the linear and saturation regimes of operation. Simulated gate leakage currents for NMOS and PMOS transistors are found to be small relative to off-state leakage current, and therefore are not considered in the SRAM yield estimation. The cell sigma, defined as the minimum number of standard deviations (for any combination of variation sources) that can result in a read failure or a write failure (*i.e.* SNM or I_w less than zero, respectively) [19], is then computed accounting for process-induced variations in device width and gate length (assuming Gaussian distributions with 3σ = 10% of nominal value) as well as RDF-induced variation in V_T (see Table 2.1).

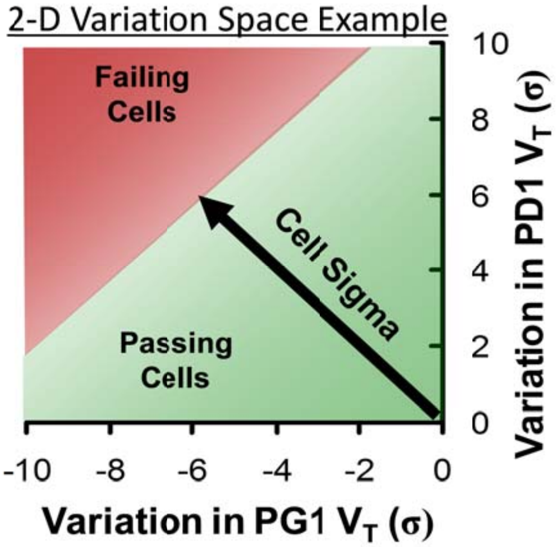

2-D Variation Space Example
Failing Cells
Passing Cells
Cell Sigma
10
8
6
4
2
0
Variation in PD1 V_T (σ)
-10
-8
-6
-4
-2
0
Variation in PG1 V_T (σ)

Fig. 2.7 Illustration of a cell sigma metric in a 2-D variation space, accounting for variation in V_T of PG and PD. The most likely failure is represented by the length of the shortest vector ("cell sigma") in the variation space emanating from the origin and terminating at the failure surface. The actual model takes into account 18 dimensions of variability, including, W, L, and V_T of the 6 transistors in the SRAM cell.

Based on published 6-T SRAM cell designs for 28 nm CMOS technology [20]-[22], the cell area is set to be 0.120 um^2 in this work. The nominal device widths (W) for pull-down (PD), pull-up (PU), and pass-gate (PG) devices are set to be 80 nm, 50 nm, and 50 nm, respectively. (Based on reports in the literature [20], 50 nm is the typical device width for this technology node.) Taking into account various layout constraints such as minimum active spacing, poly-poly spacing, poly-active overlap, contact size, and so forth, the allowable range of the active width for each transistor was estimated. Within the layout constraint of $W_{PG} + W_{PD} \leq 165$ nm, these device widths can be adjusted to tune the trade-off between read stability and write stability. The sensitivity of SNM and I_{write} vs. the width of PG are shown in Fig. 2.8. Based on the nominal design (W_{PG} = 50nm) , SSR 2 shows the highest SNM, but also lower I_{write} compared to other design cases. However, due to better trade-off between read and write stability, the width of the PG in SSR2 can be made larger such that the SNM is lowered to match that of UD's cases while increasing its write current. When the read SNM of SSR2 is matched to the UD's nominal SNM (~190 nm), it has a 21.4% higher write current as compared to the UD's cases.

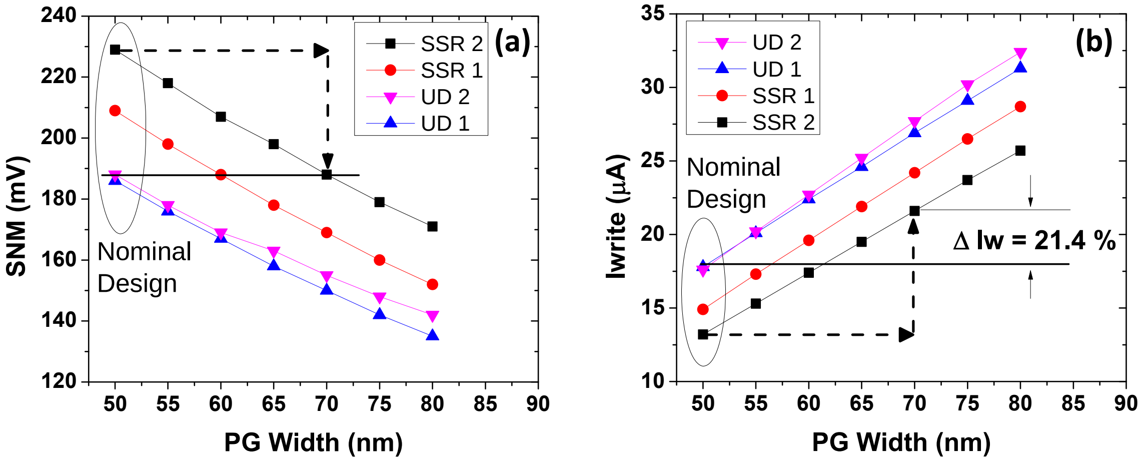


Fig. 2.8 Sensitivity of a) SNM and b) I_{write} to the width of PG as it is varied from 50 nm to 80 nm.

Cells implemented with UD devices have better writeability than read stability; therefore, W_{PD} should be increased to improve SNM and thereby maximize cell sigma. Figs. 2.9a) and 2.9b) show how the trade-off between read yield and write yield (each measured in number of cell sigmas) changes as W_{PD} is increased from 80 nm to 115 nm, for SRAM cells implemented with UD1 and UD2 devices, respectively. (Note that writeability depends primarily on the strength of the PU device relative to the PG device, so that changes in W_{PD} have little impact on I_w yield.) Cells implemented with SSR devices have better read stability than writeability; therefore, W_{PG} should be increased to improve writeability and thereby maximize cell sigma. Figs. 2.9c) and 2.9d) show how the trade-off between read yield and write yield changes as W_{PG} is increased from 50 nm to 80 nm, for SRAM cells implemented with SSR1 and SSR2 devices, respectively.

For large-capacity SRAM, six-sigma yield for both read and write operation is required. For UD1 devices, the minimum cell operating voltage ($V_{MIN,\ SRAM}$) is 0.95 V, and for UD2 devices $V_{MIN,SRAM}$ = 0.75 V. Comparing UD *vs.* SSR devices with the same background doping, it can be seen that $V_{MIN,\ SRAM}$ is reduced by ≥ 0.25 V with SSR doping. This translates to dynamic power savings of >50%. It is interesting to note that for SSR2 doping, $V_{MIN,\ SRAM}$ is limited not by SNM yield but by I_W yield; hence, further reduction in $V_{MIN,\ SRAM}$ can be achieved by improving the strength of the NMOS transistors relative to the PMOS transistors, *e.g.* by employing strain technology [20] or incorporating oxygen partial monolayers into the channel region [13] to boost electron mobility.

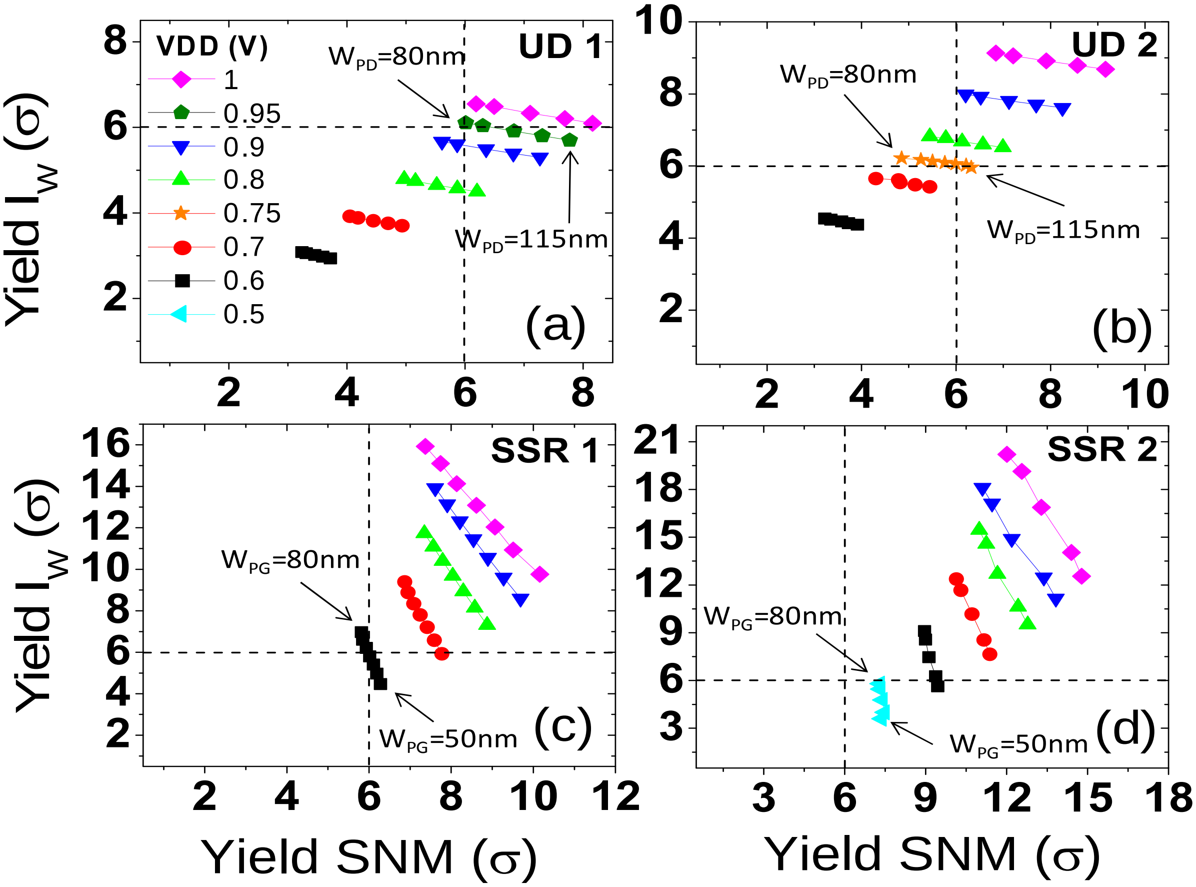


Fig. 2.9 SRAM yield result for a) UD 1, b) UD 2, c) SSR 1, and d) SSR 2. For UD, W_{PD} = 80 – 115 nm, W_{PU} = 50nm, W_{PG} = 50nm. For SSR, W_{PD} = 80 nm, W_{PU} = 50 nm, W_{PG} = 50–80 nm.

2.5 Summary

Short-channel effects can be effectively mitigated in 28 nm L_g MOSFETs by employing a super-steep retrograde (SSR) channel doping profile, albeit at the cost of degraded transistor drive current due to enhanced body factor. Estimations of six-transistor (6-T) SRAM cell yield indicate SSR doping can provide for 33% reduction in $V_{MIN,\ SRAM}$ – which translates into >50% dynamic power savings – as compared against uniform channel doping.

2.6 References

[1] D.J. Frank; R.H. Dennard; E. Nowak; P.M. Solomon; Y. Taur; H.S.P Wong, "Device scaling limits of Si MOSFETs and their application dependencies," *Proc. IEEE*, vol. 89, no. 3, Mar. 2001.

[2] A. E. Carlson, "Device and circuit techniques for reducing variation in nanoscale SRAM," Ph.D. dissertation, UC Berkeley, May 2008.

[3] C. Auth; C. Allen; A. Blattner; D. Bergstrom; M. Brazier; M. Bost; M. Buehler; V. Chikarmane; T. Ghani; T. Glassman; R. Grover; W. Han; D. Hanken; M. Hattendorf; P. Hentges; R. Heussner; J. Hicks; D. Ingerly; P. Jain; S. Jaloviar; R. James; D. Jones; J. Jopling; S. Joshi; C. Kenyon; H. Liu; R. McFadden; B. Mcintyre; J. Neirynck; C. Parker; L. Pipes; I. Post; S. Pradhan; M. Prince; S. Ramey; T. Reynolds; J. Roesler; J. Sandford; J. Seiple; P. Smith; C. Thomas; D. Towner; T. Troeger; C. Weber; P. Yashar; K. Zawadzki; K. Mistry, "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors," *VLSI Technology (VLSIT), 2012 Symposium on* , vol., no., pp.131,132, 12-14 June 2012.

[4] N. Planes; O. Weber; V. Barral; S. Haendler; D. Noblet; D. Croain; M. Bocat; P. Sassoulas; X. Federspiel; A. Cros; A. Bajolet; E. Richard; B. Dumont; P. Perreau; D. Petit; D. Golanski; C. Fenouillet-Beranger; N. Guillot; M. Rafik; V. Huard; S. Puget; X. Montagner; M.-A. Jaud; O. Rozeau; O. Saxod; F. Wacquant; F. Monsieur; D. Barge; L. Pinzelli; M. Mellier; F. Boeuf; F. Arnaud; M. Haond, "28nm FDSOI Technology Platform for High-Speed Low-Voltage Digital Applications," in *VLSI Symp. Tech. Dig.*, 2012.

[5] D.A. Antoniadis, J.E. Chung, "Physics and technology of ultra short channel MOSFET devices," in *IEDM Tech. Dig.*, 1991, pp 21-24.

[6] T. Skotnicki, "Advanced Architectures for 0.18-0.12um CMOS generations," in *Proc. ESSDERC*, 1996, pp.505-514.

[7] S.E. Thompson, P.A. Packan, M.T. Bohr, "Linear versus saturated drive current: Tradeoffs in super steep retrograde well engineering," in *VLSI Symp. Tech. Dig.*, 1996, pp. 12 -13.

[8] I. De and M. Osburn, "Impact of Super-Steep-Retrograde Channel Doping Profiles on the Performance of Scaled Devices," *IEEE Trans. Electron Devices*, vol. 46, no. 8, Aug. 1999, pp. 1711-1717.

[9] A. Asenov and S. Saini, "Suppression of random dopant-induced threshold voltage fluctuations in sub-0.1-um MOSFETs with epitaxial and δ-doped channels," *IEEE Trans. Electron Devices*, vol. 46, no. 8, Aug. 1999, pp. 1718-1724.

[10] A. Hokazono ; H. Itokawa; N. Kusunoki; I. Mizushima; S. Inaba; S. Kawanaka; Y. Toyoshima, "Steep channel & Halo profiles utilizing boron-diffusion-barrier layers (Si:C) for 32 nm node and beyond," in *VLSI Symp. Tech. Dig.*, 2008, pp. 112–113.

[11] A. Hokazono, H. Itokawa; I. Mizushima; S. Kawanaka; S. Inaba; Y. Toyoshima, "Steep channel profiles in n/pMOS controlled by borondoped Si:C layers for continual bulk-CMOS scaling," in *IEDM Tech. Dig.*,2009, pp. 673–676.

[12] R.J. Mears; N. Xu; N. Damrongplasit; H. Takeuchi; R.J. Stephenson; N.W. Cody; A. Yiptong; X. Huang; M. Hytha; T.-J. King-Liu, "Simultaneous Carrier Transport Enhancement and Variability Reduction in Si MOSFETs by Insertion of Partial Monolayers of Oxygen," in *Si Nano. elec. Workshop (SNW)*, 2012.

[13] N. Xu; N. Damrongplasit; H. Takeuchi; R.J. Stephenson; N.W. Cody; A. Yiptong; X. Huang; M. Hytha; R.J. Mears; T.-J. Liu, "MOSFET performance and scalability enhancement by insertion of oxygen layers," *Electron Devices Meeting (IEDM), 2012 IEEE International* , vol., no., pp.6.4.1,6.4.4, 10-13 Dec. 2012 .

[14] Sentaurus User's Manual, Synopsys, Inc. Mountain View, CA, 2011.

[15] N. Sano, K. Matsuzawa, M. Mukai, and N. Nakayama, "Role of long-range and short-range coulomb potentials in threshold characteristics under discrete dopants in sub-0.1 μm Si-MOSFETs," in *IEDM Tech. Dig*., 2000, pp. 275-278.

[16] H. Fuketa; T. Yasufuku; S. Iida; M. Takamiya; M. Nomura; H. Shinohara; T. Sakurai, "Device-Circuit Interactions in Extremely Low Voltage CMOS Designs (invited)," in *IEDM Tech. Dig*., 2011.

[17] E. Seevinck; F.J. List; J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *Solid-State Circuits, IEEE Journal of* , vol.22, no.5, pp.748,754, Oct 1987.

[18] C. Wann; R. Wong; D.J. Frank; R. Mann; S.-B. Ko; P. Croce; D. Lea; D. Hoyniak; Y.-M. Lee; J. Toomey; M. Weybright; J. Sudijono, "SRAM cell design for stability methodology," *Proc. IEEE VLSI-TSA Int. Symp. VLSI Technol.*, 2005, pp.21-22.

[19] C. Shin; N. Damrongplasit; X. Xin; Y. Tsukamoto; B. Nikolic; T.-J. King, "Performance and Yield Benefits of Quasi-Planar Bulk CMOS Technology for 6-

T SRAM at the 22-nm Node," *IEEE Trans. Electron Devices*, vol. 58, no. 7, July. 2011, pp. 1846-1854.

[20] C.W. Liang; M.T. Chen; J.S. Jenq; W.Y. Lien; C.C. Huang; Y.S. Lin; B.J. Tzau; W.J. Wu; Z.H. Fu; IC. Wang; P.Y. Chou; C.S. Fu; C.Y. Tzeng; K.L. Chiu; L.S. Huang; J.W. You; J.G. Hung; Z.M. Cheng; B.C. Hsu; H.Y. Wang; Y.H. Ye; J.Y. Wu; C.L. Yang; C.C. Huang; C.C. Chien; Y.R. Wang; C.C. Liu; S.F. Tzou; Y.H. Huang; C.C. Yu; J.H. Liao; C.L. Lin; D.F. Chen; S.C. Chien; IC. Chen, "A 28nm Poly/SiON CMOS Technology for Low-Power SoC Applications," in *VLSI Symp. Tech. Dig.,* 2011.

[21] F. Arnaud; A. Thean; M. Eller; M. Lipinski; Y.W. Teh; M. Ostermayr; K. Kang; N-S Kim; K. Ohuchi; J.-P. Han; D.R. Nair; J. Lian; S. Uchimura; S. Kohler; S. Miyaki; P. Ferreira; J.-H. Park; M. Hamaguchi; K. Miyashita; R. Augur; Q. Zhang; K. Strahrenberg; S. ElGhouli; J. Bonnouvrier; F. Matsuoka; R. Lindsay; J. Sudijono; F.S. Johnson; J.-H. Ku; M. Sekine; A. Steegen; R. Sampson, "Competitive and cost effective high-k based 28nm CMOS technology for low power applications," *Electron Devices Meeting (IEDM), 2009 IEEE International* , vol., no., pp.1,4, 7-9 Dec. 2009

[22] S.-Y. Wu; J.J. Liaw; C.Y. Lin; M.-C. Chiang; C.K. Yang; J.Y. Cheng; M.-H. Tsai; M.Y. Liu; P.H. Wu; C.H. Chang; L.C. Hu; C.I. Lin; H.F. Chen; S.Y. Chang; S.H. Wang; P.Y. Tong; Y.L. Hsieh; K.H. Pan; C.-H. Hsieh; C.H. Chen; C.H. Yao; C.-C. Chen; T.-L. Lee; C.W. Chang; H.J. Lin; S.C. Chen; J.H. Shieh; M.H. Tsai; S.M. Jang; K.S. Chen; Y. Ku; Y.C. See; W.J. Lo, "A highly manufacturable 28nm CMOS low power platform technology with fully functional 64Mb SRAM using dual/tripe gate oxide process," *VLSI Technology, 2009 Symposium on* , vol., no., pp.210,211, 16-18 June 2009

Chapter 3

Threshold Voltage and DIBL Variability Modeling based on Forward and Reverse Measurement for SRAM and Analog MOSFETs

3.1 Introduction

Variability in transistor threshold voltage (V_T) worsens as transistor size continues to be scaled down, severely affecting SRAM cell yield and degrading the performance of analog circuits, preventing reductions in V_{DD} [1-9]. Random sources of variability become dominant in deep-sub-micron transistor technology, and include Random Dopant Fluctuation (RDF), Work Function Variation (WFV), and Gate Line-Edge Roughness Variation (LER) [10-18]. It is imperative to have a physical understanding of how these variability sources affect the characteristic of a transistor. For example, RDF-induced variation is not only caused by the different number of dopant atoms statistically distributed among transistors, but also the position at which these atoms are located within the channel region. A detailed variability component breakdown is necessary to fully describe V_T variability and correlations.

A method to investigate the impact of RDF-induced variability by comparing the current-*vs*.-voltage (I-V) characteristics of a transistor operating in forward (F) mode *vs*. reverse (R) mode is proposed in [19]. Source-Drain positional asymmetry due to RDF has been studied in [19-24] and results in asymmetric electrical characteristics between the forward and reverse modes, but quantitative analysis with variability component breakdown has not been shown. This chapter describes a methodology to breakdown $V_{T\ LIN}$, $V_{T\ SAT}$, and non-Gaussian drain induced-barrier lowering (DIBL) variability into more fundamental terms and also predicts the bivariate correlations between V_T and DIBL, and has been validated using SRAM and analog transistor pairs [25].

Even though the MOS transistor is a symmetric device by design, there can be a significant difference between its I-V characteristics measured in forward and reverse modes, due to RDF. This difference is not systematic but rather random in nature, with no intentional difference in median V_T between F *vs*. R mode. It can have a profound impact on the operation of circuits in which transistors operate in both modes, an example being an SRAM cell. During a cell read operation, the bit lines typically are pre-charged high to V_{DD}. If the storage node is in the '0' data state, current flows from the bit line (BL) through the pass-gate (PG) and pull-down (PD) transistor stack, as shown in Fig. 3.1(a). In this case, the bit line is acting as the drain terminal and the storage node is acting as the source terminal for the PG device. During a write '0' operation, on the other hand, if the storage node is initially in the '1'data state, current flows through the pull-up (PU) and PG transistor stack down to the BL, as shown in Fig. 3.1(b). In this case, the bit line is acting as the source terminal and the storage node is acting as the drain terminal for the PG device. Therefore, the PG transistor in an SRAM cell is an example of a device which must operate in both the F and R modes. Thus, it is important to account for asymmetry in transistor performance in order to have a better assessment of the SRAM cell stability metrics.

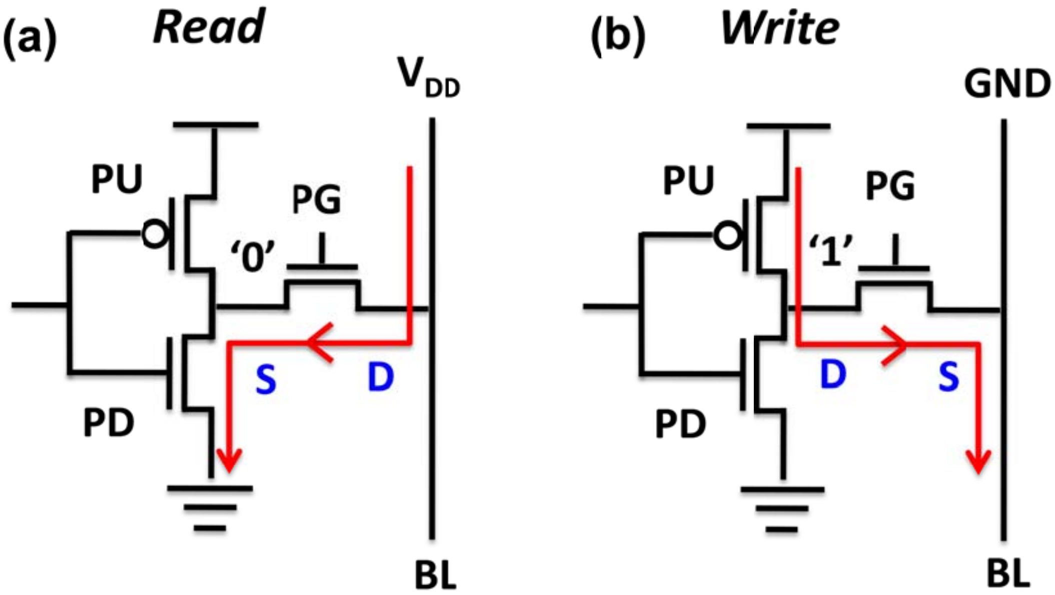


Fig. 3.1 Circuit diagrams showing the direction of current flow during (a) Read and (b) Write operation in 6T SRAM cell. The source and drain with respect to the biasing is denoted as 'S' and 'D', respectively.

From a more fundamental point of view, random asymmetry which causes differences in F vs. R operation is part of the overall device variability. It is not necessary for a transistor to operate in both F and R mode in order to be impacted by this fundamental variability. In addition, not only will the threshold voltage of a transistor be affected by such random asymmetry, the median value of DIBL and variability in DIBL also will be affected as a result. DIBL variability is important for both SRAM and analog devices. For SRAM, it is important because the PU, PD, and PG transistors operate in different regimes during read and write operation. Therefore, V_T dependence on V_{DS} bias is important and DIBL serves as a proxy for this. For analog devices, DIBL and its variability directly correlate to variability in output resistance (r_o), which in turn affects intrinsic gain and the maximum operating frequency of the device.

All of these issues emphasize the need for accurate DIBL modeling. Unfortunately, traditional Monte Carlo SPICE simulation does not take random asymmetric variation into account, and thus cannot accurately capture DIBL variability. Previously, DIBL has been reported to be non-Gaussian and a log-normal distribution was proposed to empirically fit the data [20]. The goal of this work is to explain the fundamental variability components of $V_{T\ LIN}$, $V_{T\ SAT}$, DIBL distributions as well as the correlation between these parameters using a methodical variability component decomposition procedure.

3.2 Test Structures and Modeling

3.2.1 Mismatch Pairs and Device Arrays

SRAM and analog devices fabricated using a planar 32nm High-K Metal Gate (HKMG) Partially-Depleted Silicon-On-Insulator (PDSOI) process are characterized in order to understand the underlying variability components. Fig. 3.2 shows the test structures for mismatch device pairs. The local variability of a 6T SRAM cell is studied using equivalent transistors from the left and the right halves of an SRAM cell (i.e. PG1 and PG2 form a mismatch pair). For analog devices, a mismatch pair is formed by laying out two identical transistors in close proximity.

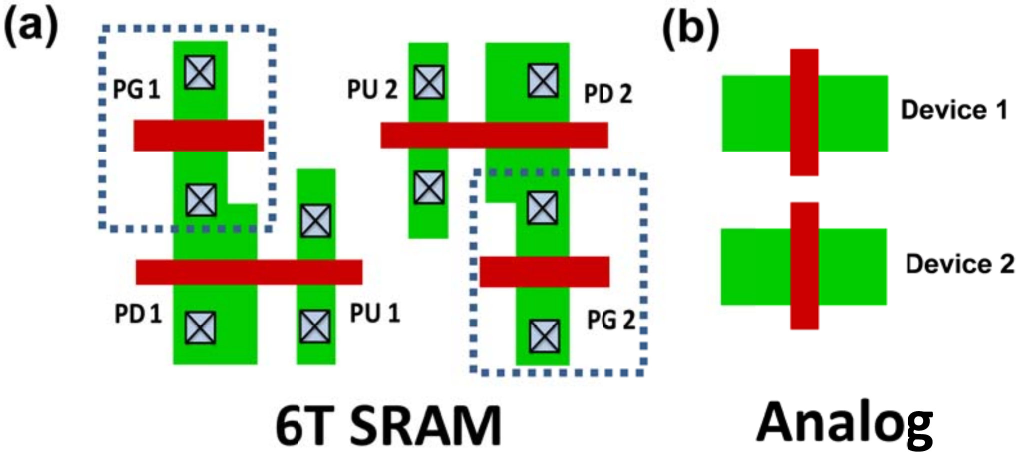


Fig. 3.2 Layouts (of active, gate and contact layers) showing (a) 6T SRAM cell consisting of individual pull-down (PD), pull-up (PU), and pass-gate (PG) transistors. (b) Analog devices consisting of identical devices are drawn in close proximity.

The variability component analysis requires that I-V characteristics for both the forward and reverse mode be measured, as illustrated in Fig. 3.3. In the forward mode, the ground terminal (GND) is connected to the source side and the supply voltage (V_{DD}) is connected to the drain side of the transistor. In a reverse mode, on the other hand, one simply interchanges the electrical biasing: GND is connected to the drain side, while V_{DD} is connected to the source side. Linear and saturation threshold voltages (at multiple V_{DD}) are collected in both modes.

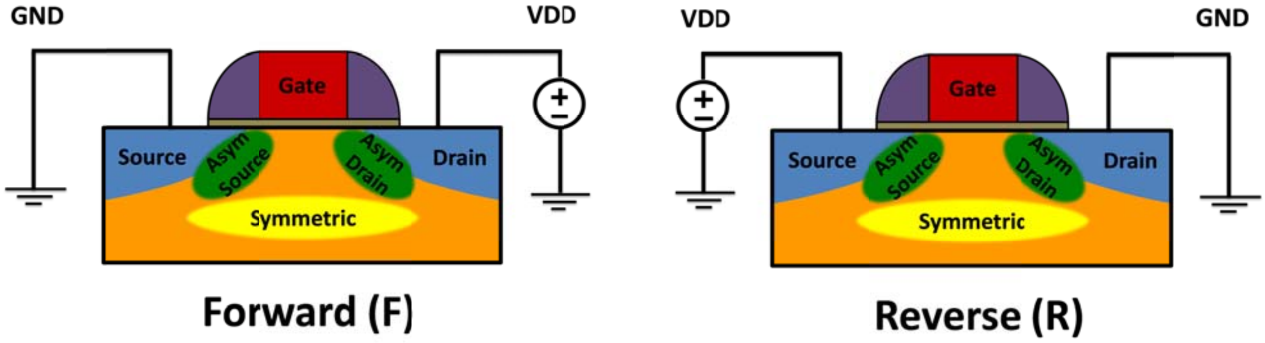


Fig. 3.3 Schematic transistor cross-sections illustrating Forward (F) vs. Reverse (R) mode biasing scheme

3.2.2 Modeling Variability Components

The overall variability is broken down into four main components: one global component (Chip Mean CM [σ_{CM}]), and three local components (Symmetric [σ_{Sym}], Asymmetric [σ_{Asym}], and L_{Eff} variation [σ_{SCE_Leff}]). Chip mean variability results from lot-to-lot, wafer-to-wafer, die-to-die (across wafer) variation, and it impacts both devices in a mismatch pair equally. Symmetric, asymmetric, and L_{Eff} variations constitute the random or local variations in the device. The symmetric component is defined as that which contributes equally to the V_T between the forward and reverse measurement of the same device under test. Sources of this type of variation are gate work function variation, and random dopant fluctuations in the well doping profile or in the V_T adjustment doping profile implanted uniformly across the entire channel region. In contrast, the asymmetric component is defined as that which can result in a V_T difference between F and R modes. The most likely source of this type of variation is random dopant fluctuations in the halo doping profiles at the source and drain sides of the device, or in the source/drain extension doping profiles. Lastly, V_T variation due to L_{Eff} is also considered. This is the electrostatic variability component that is associated with the short channel effect (SCE), which depends on the drain bias V_{DS} since SCE worsens as drain bias increases. These fundamental variability components which serve as the input to the model each are assumed to have a Gaussian distribution and to be independent of one another. Nevertheless, no assumption is made *a priori* about the distribution of the outputs of the model, including $V_{T\ LIN}$, $V_{T\ SAT,}$ and DIBL.

3.2.3 Effect of Asymmetric Variation

Halo or pocket implants are commonly used to suppress the short channel effect. However, as some have reported in literature previously, this can result in random asymmetry of the F vs. R characteristic of a transistor [19,22]. To understand the physical reasoning behind this experimental observation, it is informative to examine the conduction band edge profile along the channel region of the device, as depicted in Fig. 3.4. Let's first focus on the forward mode operation in Fig 3.4(a). The black trace indicates the profile in the linear regime (V_{DS} = 50 mV) and the red trace indicates the profile in the saturation regime (V_{DS} = 1V). For the device shown here in the linear regime, the source-side potential barrier is higher with respect to its drain-side potential. It is likely that this device happens to either have a higher halo or a lighter extension doping concentration near the source-side, and hence the maximum channel potential is located near the source in the linear regime under forward mode. Furthermore, the V_T of the device can be qualitatively associated with the height of the source-channel barrier to carrier injection, and therefore for the device shown here, the source-side barrier will determine the linear threshold voltage under forward mode ($V_{T\ LIN\ F}$).

In the saturation regime (shown by the red curve), due to the high V_{DS} being applied, the potential near the drain-side is lowered significantly, causing the maximum potential barrier along the channel to *always* appear on the source-side with respect to the mode that is being measured. This barrier height will qualitatively set the saturation threshold voltage under forward mode ($V_{T\ SAT\ F}$). It is interesting to note that the difference in the potential barrier height between linear and saturation regime qualitatively determines the drain-induced barrier lowering (DIBL $_F$) of the device in forward mode.

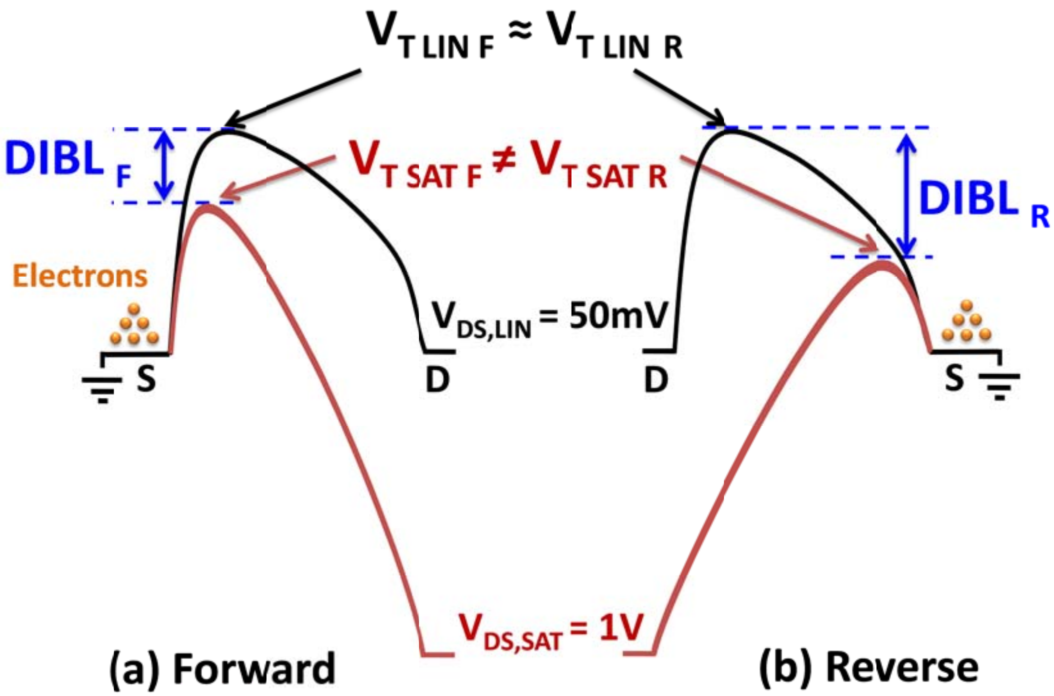


Fig. 3.4 Qualitative view of the conduction band-edge profile across the source/channel/drain regions of a MOSFET, in the linear and saturation regime of a) Forward and b) Reverse mode operation. S and D denote the source and the drain regions, respectively.

When this same device is measured in the reverse mode as shown in Fig. 3.4(b), in the linear regime the maximum source-channel potential barrier is unchanged relative to the forward mode, *i.e.* it is now located on the drain side with respect to the reverse mode measurement. This results in $V_{T\ LIN}$ being approximately the same in both F and R modes ($V_{T\ LIN\ F} \approx V_{T\ LIN\ R}$). In the saturation regime, the applied V_{DS} is large enough such that the band-edge profile near the drain side is lowered significantly, leaving the maximum potential to be always located on the source-side of the device. Thus, in reverse mode, the saturation threshold voltage ($V_{T\ SAT\ R}$) will be set by the potential barrier at the source-side which is lower (due to lower net doping in this case). The height difference between the maximum potential barrier in the linear and saturation regime is again representative of DIBL in the reverse mode (DIBL $_{R}$). It is imperative to note that for the reverse mode measurement shown here the maximum potential barrier in the channel changes from the drain-side in the linear regime to the source-side in the saturation regime. In essence, $V_{T\ LIN}$ is unchanged in F-R, whereas $V_{T\ SAT}$ can be quite different in F-R. Consequently, there can be a large difference between $V_{T\ LIN\ R}$ and $V_{T\ SAT\ R}$ for this particular device as compared to the forward mode. This implies that both the median and the standard deviation value of $DIBL_R$ will also be larger as compared to those of $DIBL_F$. Thus, random device asymmetry contributes to the differences between $V_{T\ SAT\ F}$ and $V_{T\ SAT\ R}$, and determines the degree of correlation between linear and saturation threshold voltage, which is captured in the DIBL variability metric.

3.2.4 Parameter Extraction Flow

Fig. 3.5 illustrates the parameter extraction flow for the proposed model. Threshold voltages in both the linear and saturation regimes of operation, as well as for forward and reverse modes, are extracted using a constant-current definition: 300 nA *(W/L) for NMOS; 70 nA*(W/L) for PMOS. Once experimental data is acquired, quantities such as DIBL and V_T mismatch ($V_{T\ LIN/SAT\ MM}$) can be derived with respect to different mode of measurements. Using the proposed governing equations, the four fundamental variation components mentioned earlier can be extracted. Lastly, by employing these variation components in the model, $V_{T\ LIN}$, $V_{T\ SAT}$, and DIBL are reconstructed and compared against silicon data for variability component validation.

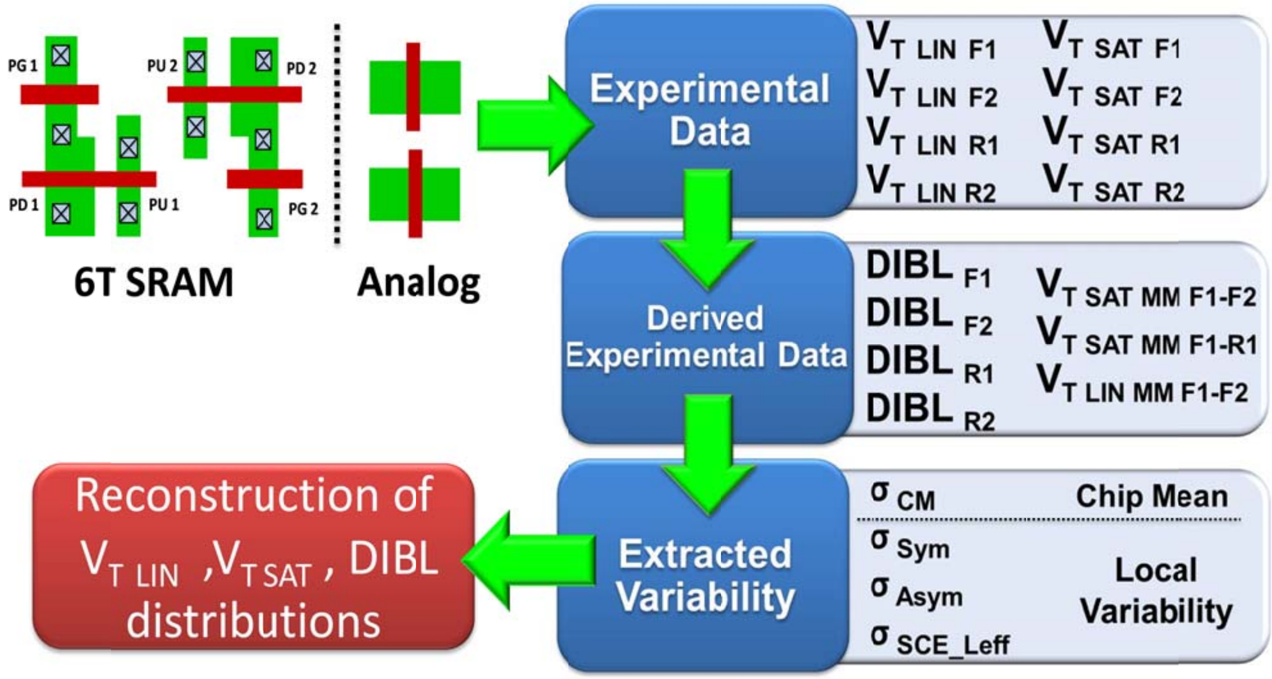


Fig. 3.5 Parameter Extraction flow

TABLE 3.1

LIST OF EQUATIONS IN VARIABILITY MODELING

(a)

$$\mathrm{DIBL}_{F/R,1/2} = V_{T\,LIN\,F/R\,1/2} - V_{T\,SAT\,F/R\,1/2}$$

$$V_{T\,SAT\,MM\,[1-2]} = V_{T\,SAT\,F/R\,1} - V_{T\,SAT\,F/R\,2}$$

$$V_{T\,SAT\,MM\,[F-R]} = V_{T\,SAT\,F\,1/2} - V_{T\,SAT\,R\,1/2}$$

$$V_{T\,LIN\,MM\,[1-2]} = V_{T\,LIN\,F/R\,1} - V_{T\,LIN\,F/R\,2}$$

(b)

$$\begin{aligned} V_{T\,LIN\,F\,1/2} = {} & V_{T0} - V_{DS,LIN} \cdot \mathrm{NDIBL} + \Delta V_{T\,CM} + \Delta V_{T\,Sym\,1/2} \\ & + \max(\Delta V_{T\,Asym\text{-}Source\,1/2}\,,\Delta V_{T\,Asym\text{-}Drain\,1/2}) - V_{DS,LIN} \cdot \Delta V_{T\,SCE\_Leff\,F\,1/2} \end{aligned}$$

$$\begin{aligned} V_{T\,LIN\,R\,1/2} = {} & V_{T0} - V_{DS,LIN} \cdot \mathrm{NDIBL} + \Delta V_{T\,CM} + \Delta V_{T\,Sym\,1/2} \\ & + \max(\Delta V_{T\,Asym\text{-}Drain\,1/2}\,,\Delta V_{T\,Asym\text{-}Source\,1/2}) - V_{DS,LIN} \cdot \Delta V_{T\,SCE\_Leff\,R\,1/2} \end{aligned}$$

$$\begin{aligned} V_{T\,SAT\,F\,1/2} = {} & V_{T0} - V_{DS,SAT} \cdot \mathrm{NDIBL} + \Delta V_{T\,CM} + \Delta V_{T\,Sym\,1/2} \\ & + \Delta V_{T\,Asym\text{-}Source\,1/2} - V_{DS,SAT} \cdot \Delta V_{T\,SCE\_Leff\,F\,1/2} \end{aligned}$$

$$\begin{aligned} V_{T\,SAT\,R\,1/2} = {} & V_{T0} - V_{DS,SAT} \cdot \mathrm{NDIBL} + \Delta V_{T\,CM} + \Delta V_{T\,Sym\,1/2} \\ & + \Delta V_{T\,Asym\text{-}Drain\,1/2} - V_{DS,SAT} \cdot \Delta V_{T\,SCE\_Leff\,R\,1/2} \end{aligned}$$

NOTE: V_{T0} **and NDIBL are fitting parameters**

$\Delta V_{T\,\text{'Extracted Variability'}} \equiv$ **Gaussian Random Variable** $N(0,\sigma^2_{\text{'Extracted Variability'}})$

(c)

$$\sigma_{Asym} = \sigma_{Asym-Source\,1/2} = \sigma_{Asym-Drain\,1/2}$$

$$\sigma_{Sym} = \sigma_{Sym\,1/2}$$

$$\sigma_{SCE\_Leff} = \sigma_{SCE\_Leff\quad F/R\,1/2}$$

$$\sigma_{MAX(\Delta V_{T\,\,Asym-Source/Drain\,1/2})} \approx \frac{\sigma_{Asym}}{\sqrt[4]{2}}$$

(d)

$$\sigma^2_{V_T\,SAT\,MM[F-R]} = 2 \cdot \sigma^2_{Asym} + 2 \cdot (V_{DS,SAT})^2 \cdot \sigma^2_{SCE\_Leff}$$

$$\sigma^2_{V_T\,SAT\,MM[1-2]} = 2 \cdot \sigma^2_{Sym} + 2 \cdot \sigma^2_{Asym} + 2 \cdot (V_{DS,SAT})^2 \cdot \sigma^2_{SCE\_Leff}$$

$$\sigma^2_{V_T\,LIN\,MM[1-2]} = 2 \cdot \sigma^2_{Sym} + 2 \cdot \frac{\sigma^2_{Asym}}{\sqrt{2}} + 2 \cdot (V_{DS,LIN})^2 \cdot \sigma^2_{SCE\_Leff}$$

$$\sigma^2_{CM} = \sigma^2_{V_T} - \sigma^2_{Loc}\,, where \quad \sigma_{Loc} = \frac{\sigma_{V_T\,LIN\,MM[1-2]}}{\sqrt{2}}$$

Table 3.1(a) summarizes the equations used to obtain the derived experimental quantities. Note that the derived DIBL value presented here is the absolute DIBL value (i.e. the difference between $V_{T\ LIN}$ and $V_{T\ SAT}$, without normalizing to the difference in V_{DS} bias used in linear and saturation regimes). Table 3.1(b) summarizes how the linear and saturation threshold voltages for forward and reverse modes are derived using fundamental V_T components, namely chip mean (CM), symmetric (Sym), asymmetric (Asym), and short channel effect due to L_{Eff} (SCE_L_{eff}). Focusing first on the modeling of a linear threshold voltage of a forward mode device ($V_{T\ LIN\ F}$), the equation consists of two fitting parameters: V_{T0} and NDIBL (Normalized DIBL). V_{T0} and NDIBL are used to adjust for the mean of the threshold voltage distribution. It is important to note that these two fitting parameters are simply constant; they do not in any way affect the calculation of the standard deviation of V_T. In addition to the two fitting parameters, the threshold voltage also comprises a random component pertaining to four fundamental variability components (i.e. $\Delta V_{T\ CM}$, $\Delta V_{T\ Sym}$, $\Delta V_{T\ Asym}$, $\Delta V_{T\ SCE\_Leff}$) . Each component of V_T variability is assumed to have a Gaussian distribution, with a mean value of zero and a standard deviation corresponding to that of the extracted variability components. Additionally, in the linear regime of operation, since only the maximum potential barrier matters, the asymmetric variation component captures this physical effect mathematically through the 'max' function. The short channel effect is captured through the L_{Eff} component, which is a function of drain bias. Its effect is to lower the nominal V_T, as indicated by the negative sign. $V_{T\ LIN\ R}$ can be constructed in the same manner as $V_{T\ LIN\ F}$.

For $V_{T\ SAT}$, the main difference is that the potential barrier will always appear at the source-side due to the large V_{DS} bias, which pulls down the electron conduction band energy near the drain-side. As a result, the source-side barrier will always determine the saturation threshold voltage for $V_{T\ SAT}$; the potential on the drain-side has little influence on the overall V_T. Therefore, only the asymmetric variation component at the source-side is taken into account in the modeling. For $V_{T\ SAT}$ in forward mode, the asymmetric variation at the source-side is denoted as $\Delta V_{T\ Asym\text{-}Source}$. But for $V_{T\ SAT}$ in the reverse mode, the source-side is actually what was previously denoted as the drain-side in the forward mode. Hence the asymmetric variation at the drain-side, $\Delta V_{T\ Asym\text{-}Drain}$, is included in the modeling of $V_{T\ SAT\ R}$ instead.

The modeling assumptions are summarized in Table 3.1(c). The standard deviation of variation of the asymmetric component is assumed to be the same for both devices in a mismatch pair, and for forward *vs.* reverse modes. The symmetric component of variation is also assumed to have the same standard deviation for both devices in a mismatch pair. The standard deviation of variation due to L_{Eff} is assumed to be the same for both devices in a mismatch pair, and for forward *vs.* reverse modes. There is no closed-form expression for the standard deviation of the max function between two Gaussian random variables. Thus, the standard deviation due to the max function between the source-side and drain-side asymmetric variation can be well approximated (as validated by Monte Carlo simulation) by the standard deviation of the asymmetric variation divided by the fourth root of 2.

Table 3.1(d) describes the governing equations used to extract the fundamental variability components from standard deviations of the derived experimental data. These were derived using the V_T equations in Table 3.1(b). From experimental measurements the standard deviations for $V_{T\ SAT\ MM\ [F\text{-}R}V_{T\ SAT\ MM\ [1\text{-}2]}$, and $V_{T\ LIN\ MM[1\text{-}2]}$ are determined. Then, using the equations in Table

3.1(d) one can solve for the three unknowns: σ_{Sym}, σ_{Asym}, and $\sigma_{SCE,Leff}$. The chip mean component of variation σ_{CM} can be calculated directly from the experimental data by assuming that random variation and systematic variation components are uncorrelated, and local variations between the devices in a mismatch pair are uncorrelated [12]. Once the four fundamental variability components are extracted, the entire set of $V_{T\ LIN}$ and $V_{T\ SAT}$ distributions including the median and standard deviation can be reconstructed and validated against silicon data.

3.3 Results

3.3.1 Variability Component Analysis in SRAM

The histograms of $V_{T\ LIN}$, $V_{T\ SAT}$, and DIBL for PG transistors is shown in Fig. 3.6. The number of mismatch pairs used in this study is 1900. Experimental distributions are shown in red while the modeling results are shown in black. The distributions of $V_{T\ LIN}$ and $V_{T\ SAT}$ are found to be well approximated by Gaussian functions. On the other hand, DIBL distribution is experimentally found to be non-Gaussian and the model is able to capture the non-Gaussian nature of this distribution with good accuracy. This observation is consistent with the published literature in which it was found that the DIBL distribution is better described by a Log-Normal distribution [20]. The key point to note here is that non-Gaussian behavior of DIBL can be reproduced by the model without having to assume any empirical distribution for DIBL *a priori*.

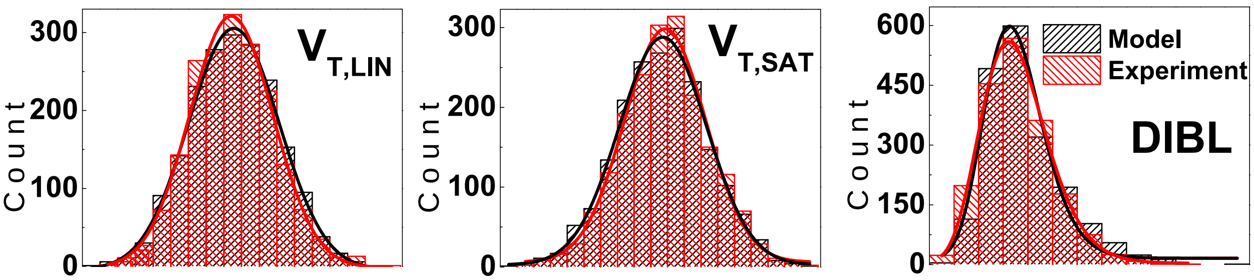


Fig. 3.6 Distributions of (a) $V_{T\ LIN}$ [@V_{DS}=50mV] (b) $V_{T\ SAT}$ [@V_{DS}=1V] (c) DIBL. $V_{T\ LIN}$ and $V_{T\ SAT}$ have nearly Gaussian distributions, but DIBL clearly does not follow a normal distribution.

Fig. 3.7 shows a correlation plot of threshold voltage values for F *vs*. R mode operation of the same device. $V_{T\ LIN}$ is shown in red (model) and blue (experiment) symbols. This result is also consistent with [20]; linear V_T values for forward and reverse modes of the same device are almost identical as expected from the aforementioned analysis of the conduction band-edge profile. For $V_{T\ SAT}$ shown in black (model) and green (experiment) symbols, there can be much larger differences between F *vs*. R mode values. The primary reason for this weaker correlation between $V_{T\ SAT\ F}$ and $V_{T\ SAT\ R}$ is random asymmetric variability and, to a lesser extent, the contribution of the L_{Eff} component since its effect also increases with increasing drain bias.

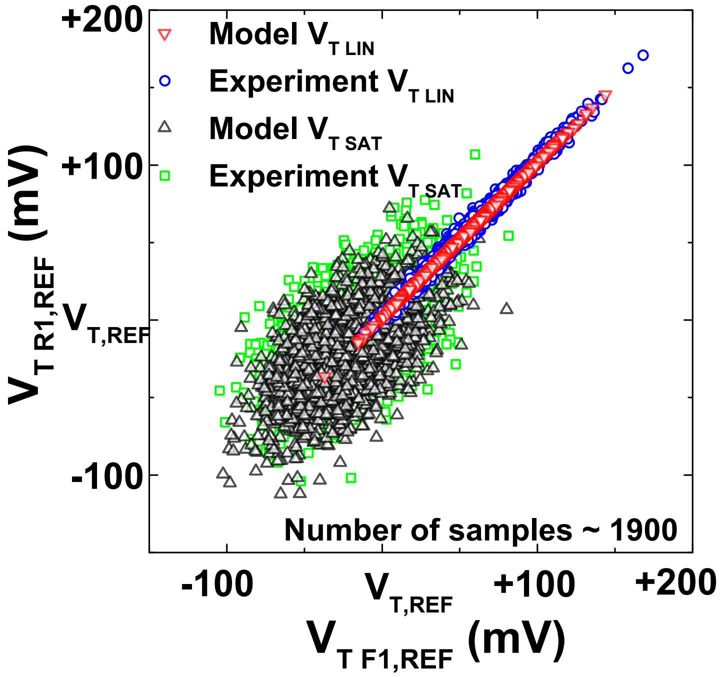


Fig. 3.7 Strong correlation between F vs. R mode is seen for $V_{T\ LIN}$. Weaker correlation is seen for $V_{T\ SAT}$.

A device with a higher $V_{T\ SAT}$ in the forward mode would likely have a maximum potential barrier near the source-side in the linear regime of operation. Therefore, a device with a higher $V_{T\ SAT}$ in the forward mode is denoted as a "source-limited" device. A similar argument can be made for a device that has a higher $V_{T\ SAT}$ in the reverse mode. In this case, it is likely that this particular device has a maximum potential barrier near the drain-side (with respect to the forward mode) in the linear regime of operation, and hence it is denoted as a "drain-limited" device. Consequently, the ensemble of the devices can be broken down into two groups, depending on where the maximum potential barrier along the channel is located in the linear regime of operation.

The device ensemble is sorted into source-limited and drain-limited devices. Based on this grouping, the median threshold voltage of the source-limited and drain-limited devices can be calculated separately. Fig. 3.8 shows the experimental data of median threshold voltage as a function of drain bias V_{DS}, for source-limited devices (solid black curve), drain-limited devices (dotted blue curve), and all devices (red dashed curve).

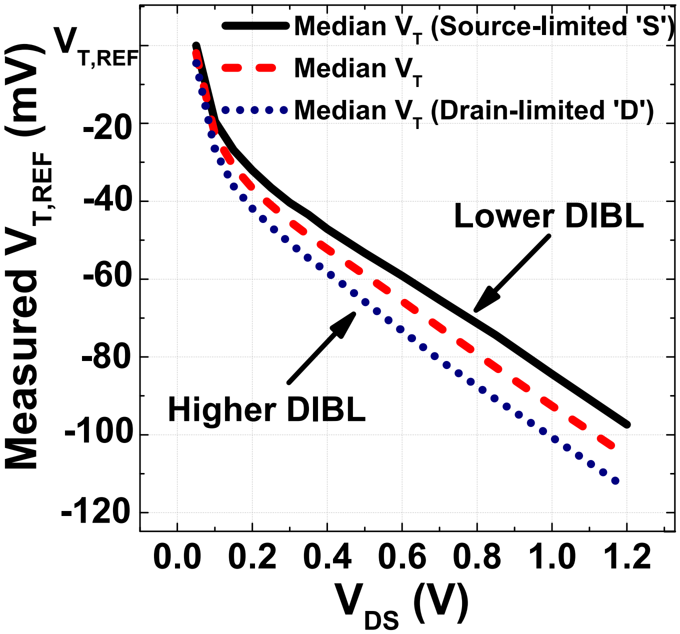


Fig. 3.8 Median value of threshold voltage as a function of drain voltage. Source-limited devices show lower DIBL compared to drain-side devices.

At low V_{DS}, the median V_T of the source- and drain-limited populations are the same. But as V_{DS} increases, the median V_T for source-limited devices does not decrease as rapidly, implying that DIBL is lower for these devices as compared to drain-limited devices. The median V_T for the entire population of devices is in fact an average of the source- and drain-limited devices' median V_T values.

The experimental result of σV_T mismatch ($\sigma V_{T\ MM\ 1\text{-}2}$) as a function of V_{DS} is shown in Fig. 3.9. The overall V_T mismatch variation is represented by the black curve. As shown, the V_T mismatch variation increases with increasing drain bias. To explain this effect, it is informative to investigate the random asymmetric variation component of σV_T mismatch as a function of drain bias. This can be experimentally determined by measuring σV_T mismatch for forward *vs.* reverse mode ($\sigma V_{T\ MM\ F\text{-}R}$) as a function of V_{DS}. The result is plotted with a dotted blue curve in Fig. 3.9, and it is clear that mismatch variation due to asymmetric variation increases substantially with increasing drain bias. Furthermore, one can analytically remove this asymmetric variation component by measuring the variation in mismatch between two source-limited devices (or drain-limited devices) in a device pair. With the asymmetric variation component removed, σV_T mismatch does not depend on V_{DS} as shown by the red dotted line. In other words, the short channel effect component only has a small contribution to the increase in variation of V_T mismatch with increasing V_{DS}. Rather, random asymmetric variation is the major component responsible for the increase in variation of V_T mismatch with increasing drain bias, resulting in the experimental observation that $\sigma V_{T\ MM\ SAT} > \sigma V_{T\ MM\ LIN}$.

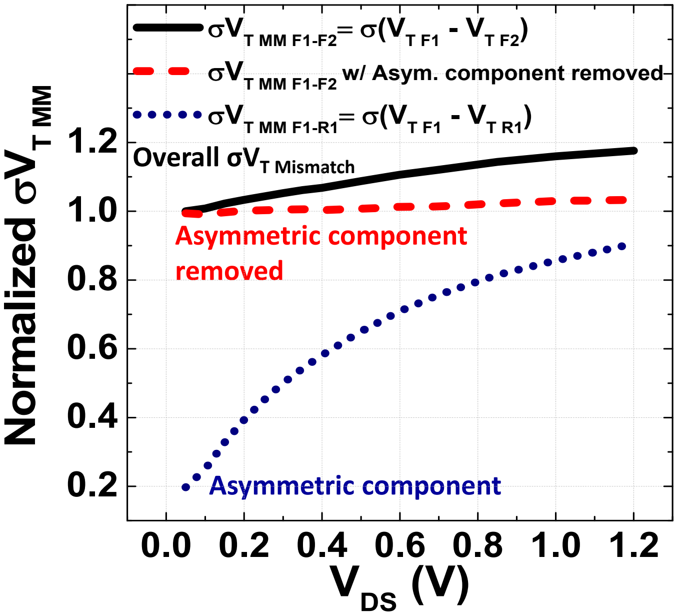


Fig. 3.9 Dependence of measured $\sigma V_{T\ MM}$ on V_{DS}. Random asymmetry variation causes $\sigma V_{T\ MM}$ to increase with increasing V_{DS}.

Correlation between linear and saturation V_T of the same device measured under a given mode (i.e. forward mode) is plotted in Fig. 3.10. The experimental data is indicated by the black symbols. Overlaid on top are the modeling results which are decomposed into two groups: the source-limited devices as shown in green and the drain-limited devices as shown in blue. Overall, there is a positive correlation between $V_{T\ SAT}$ and $V_{T\ LIN}$. By dividing the population into two groups, the model reveals why the correlation has this particular shape. The source-limited device exhibits two characteristics: 1) For a given $V_{T\ LIN}$, the mean of the corresponding $V_{T\ SAT}$ of source-limited device shows a larger value as compared to drain-limited device (lower median DIBL) 2) Correlation between linear and saturation threshold voltage for source-limited device is much greater compared to drain-limited devices (lower DIBL variability). The reason behind these observations can be explained by recalling the conduction band-edge profile introduced previously in Fig 3.4. Source-limited devices show higher correlation between $V_{T\ SAT}$ and $V_{T\ LIN}$ because the maximum potential appears near the source-side in both the linear and saturation regime. On the other hand, for the drain-limited devices, the maximum potential in the channel shifts from near the drain-side in the linear regime to near the source-side in the saturation regime, resulting in lower correlation between $V_{T\ LIN}$ and $V_{T\ SAT}$. Therefore, random asymmetric variation which introduces asymmetry in the potential profile will result in wider distributions of $V_{T\ SAT}$ and $V_{T\ LIN}$, as is evident for the drain-limited devices. The overall distribution of $V_{T\ SAT}$ vs. $V_{T\ LIN}$ is a combination of distributions for these two subsets.

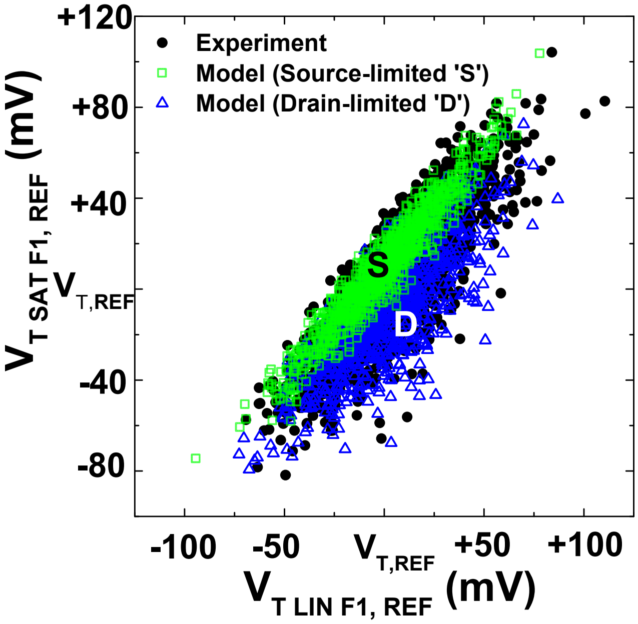


Fig. 3.10 Correlation between $V_{T\ LIN\ F1}$ and $V_{T\ SAT\ F1}$. The distribution is broken down into source-limited (S) and drain-limited (D) devices.

The relationship between DIBL and V_T can also be understood by comparing the modeling result to the experimental data. In Fig. 3.11(a), forward DIBL ($V_{T\ LIN\ F1} - V_{T\ SAT\ F1}$) is plotted against linear threshold voltage $V_{T\ LIN}$. The variation and median of $V_{T\ LIN}$ of the source- and drain- limited devices are essentially the same (i.e. this is seen by the same horizontal spread in the scatter plot). However, there is a large difference in the corresponding DIBL value. In particular, the drain-limited devices show a higher median DIBL and a much larger variability. Again, this is because drain-limited devices are prone to random device asymmetry, resulting in a weak correlation between linear and saturation threshold voltage. As a result, this helps explain the origin of the 'dome' shape when plotting DIBL vs. $V_{T\ LIN}$ that has been reported elsewhere in literature. Similarly, a plot between DIBL vs. $V_{T\ SAT}$ can also be constructed. As expected, one observes an anti-correlation relationship since higher $V_{T\ SAT}$ implies lower DIBL. Moreover, the result suggests that drain-limited devices are responsible not only for the increase in DIBL value, but also the increase in DIBL variability. We can conclude that DIBL and its variability is not only electrostatic in nature but is also impacted by components of RDF.

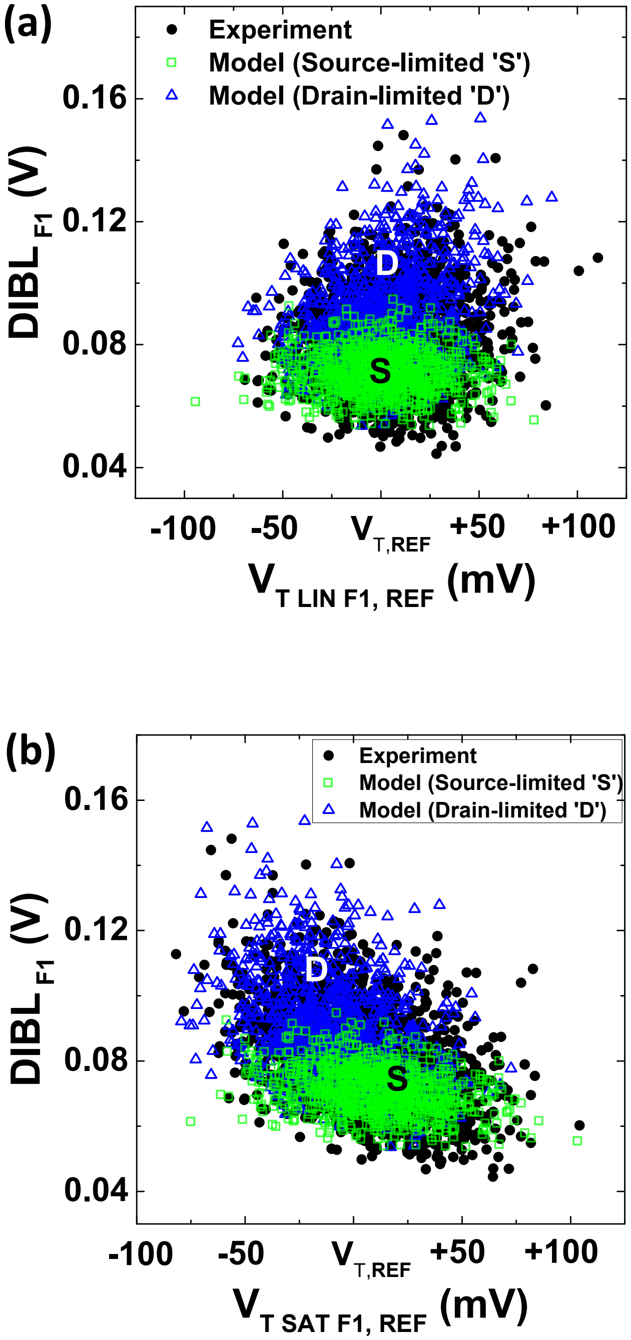


Fig. 3.11 (a) The weak correlation between $V_{T\ LIN\ F1}$ and $DIBL_{F1}$ is comprised of two populations (S and D), driven by random positional asymmetry. (b) Anti-correlation between $V_{T\ SAT\ F1}$ and $DIBL_{F1}$. Drain-limited devices increase the mean of DIBL and have larger DIBL variability.

It is also informative to examine the relationship between DIBL in the forward mode vs. DIBL in the reverse mode as shown in Fig. 3.12. The result reveals a weak anti-correlation shape between $DIBL_F$ and $DIBL_R$. The model is able to explain why the correlation has this particular characteristic by breaking down the entire population into source- and drain- limited devices. Focusing first on the source-limited devices, the mean value of the forward DIBL is generally lower compared to the reverse DIBL, as illustrated by the vertical ellipsoidal shape. This traces back to the fact that source-limited device has maximum potential barrier closer to the source-side in the linear regime. Therefore, DIBL measured in forward mode is better suppressed as compared to measuring it in the reverse mode. The same argument applies for the drain-limited device. Since the maximum potential barrier is located at the drain end in the linear regime, measuring in reverse mode ensures that the maximum potential barrier does not shift position as device is biased from linear to saturation regime, and thus DIBL for reverse mode will be smaller compared to DIBL for forward mode. The superposition of these two distribution results in the particular anti-correlation seen in the experimental data.

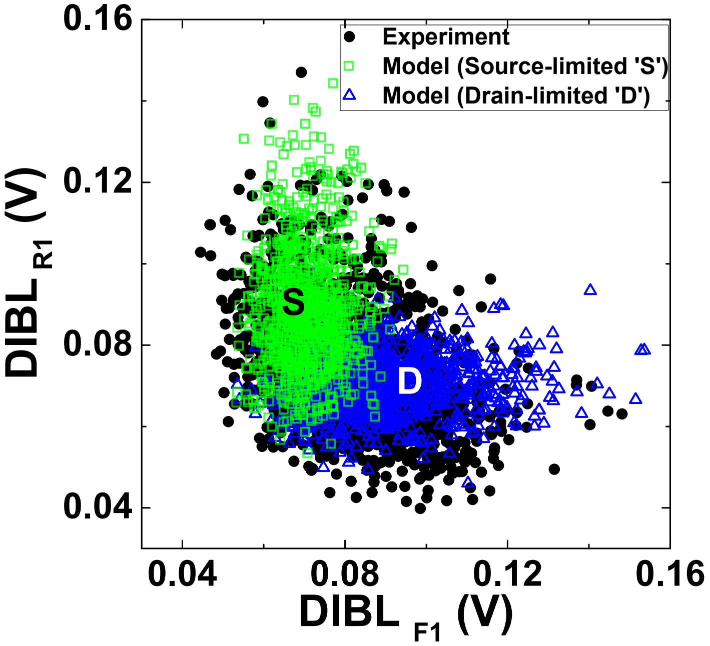


Fig. 3.12 Anti-correlation between forward and reverse DIBL mismatch ($DIBL_{F1}$ vs. $DIBL_{R1}$) is driven by random positional asymmetry and is comprised of source- and drain-limited devices.

In addition to pass-gate devices in the SRAM cell, the modeling was also done for pull-up and pull-down devices. Similar trends and correlations were observed for these devices as well.

3.3.2 Variability Component Analysis in Analog Devices

The same model that is used to describe variability in SRAM devices was applied to analog devices as well. Experimental data for analog devices was analyzed for two datasets:

short gate length and long gate length devices. Sample results for an analog device with short gate length are shown in Fig. 3.13. The number of devices under test for short gate length is 4920 mismatch pairs. Excellent agreement between the silicon data and modeling results is observed. Similar to the SRAM results, high correlation exists between $V_{T\ LIN}$ values and a weak correlation exists between $V_{T\ SAT}$ values as shown in Fig 3.13 (a). The overall data can be broken down into source- and drain-limited devices, as illustrated in the correlation plot between $DIBL_F$ and $V_{T\ SAT\ F}$ in Fig. 3.13(b). The results for long gate length analog devices are shown in Fig. 3.14(a)-(b), showing similar trends.

In short, the modeling presented herein has been validated across multiple geometries and device flavors, showcasing the robustness of the component breakdown framework.

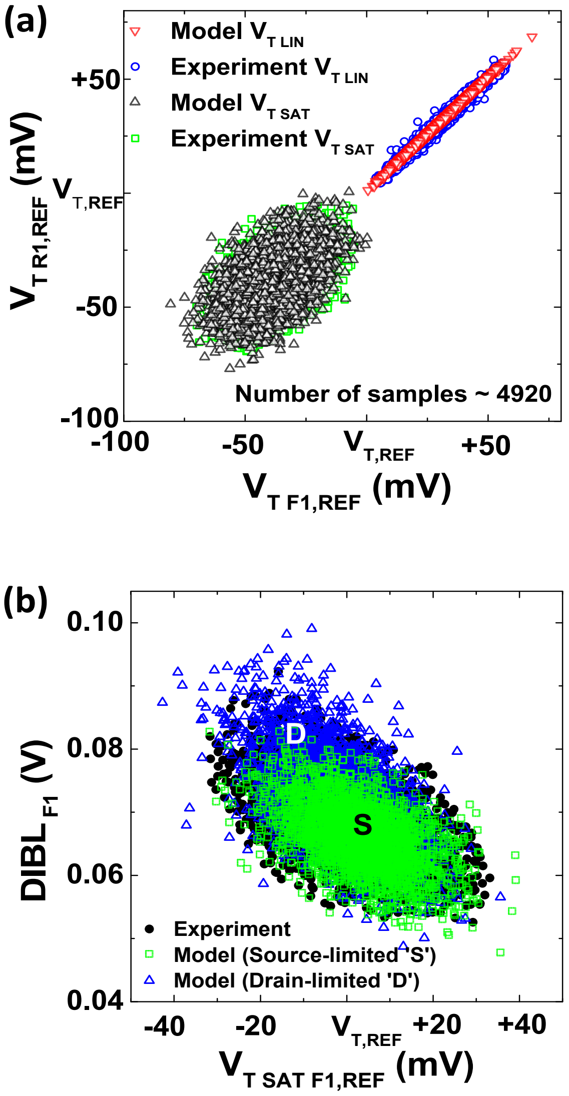


Fig. 3.13 Analog devices with short gate length L_g. a) Correlation plot between forward and reverse mode V_T b) Correlation plot of DIBL vs. $V_{T\ SAT}$.

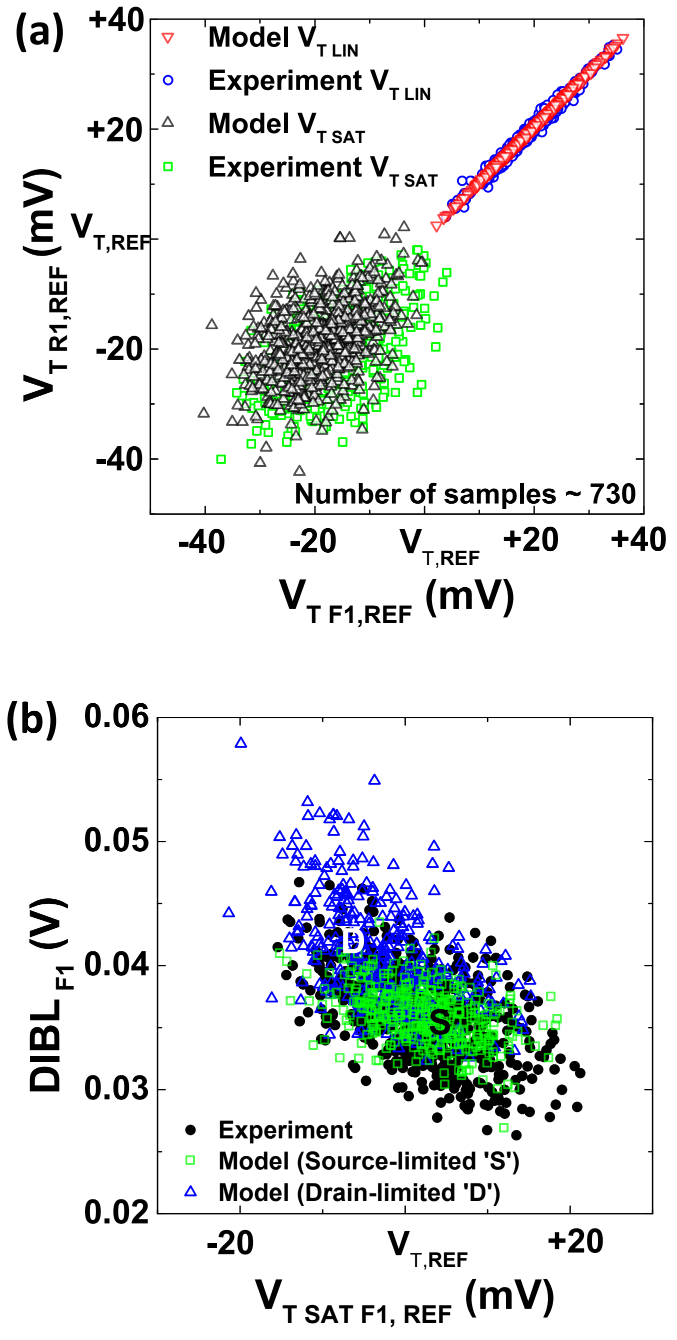


Fig. 3.14 Analog devices with long gate length L_g. a) Correlation plot between forward and reverse mode V_T b) Correlation plot of DIBL vs. $V_{T\,SAT}$.

3.3.3 Symmetric and Asymmetric Random Variability Implications for Semiconductor Process and Gate Length Scaling

The ability to decompose variability into different components not only allows a better understanding of device variability, but also provides opportunity to refine the device fabrication process. For example, when analyzing a particular flavor of transistor, if it is found that the asymmetric variation component is much larger compared to the symmetric component, then it is logical to assume that the overall variation of such device is dominated by asymmetric variation. From a processing standpoint, focus should be made on optimizing the halo doping, or the source/drain extension doping steps in the process flow. If, on the other hand, the symmetric variation component is the main contributor to V_T variation, then the emphasis should be on optimizing the well doping, or the gate stack deposition. Thus, the model allows a feedback path between device characterization and process optimization to improve transistor performance and reduce variability. Such close interaction between device and process design is even more critical for the aggressively scaled transistors in state-of-the-art CMOS technology.

It is well known that the variation in V_T mismatch increases as the geometry of a transistor reduces. Specifically, Pelgrom inferred that the variation would scale inversely with $\sqrt{W \cdot L}$ [26-27]. Alternatively, one can conclude that as the width and length of the channel region become larger, variability should become smaller. This is true in general. However, the asymmetric variability does not diminish commensurately as gate length increases (relevant for analog devices whose L_g is much larger compared to logic). As a result, DIBL and DIBL variability will not scale down well when the gate length is made to be large, contributing to poor device mismatch and large variation in the output conductance of the analog device. Moreover, this observation is also evident when analyzing a Pelgrom plot. Due to the non-scalable component of asymmetric variation with respect to gate length, one can experimentally observe a non-zero crossing point in the Pelgrom plot.

3.4 Conclusion

Variations in V_T and DIBL and their correlations can be well-captured for SRAM and analog devices by incorporating random asymmetry manifested through the difference between forward- and reverse-mode characteristics of a MOS transistor. Modeling the effects of random asymmetric variation provides a more accurate understanding of V_T and DIBL variability, and their correlations, enabling better parametric yield estimation in SRAMs and R_{OUT} variability in analog devices. We can conclude that DIBL and its variability is not only electrostatic in nature but is also impacted by components of RDF. By using this variability analysis and component breakdown modeling framework, the overall V_T and DIBL variability can be identified and then optimized by minimizing symmetric and asymmetric variation components.

3.5 References

[1] G. Zheng, A. Carlson, L.-T. Pang; K.T. Duong, T.-J. King Liu, B. Nikolic, “Large-Scale SRAM Variability Characterization in 45 nm CMOS,” *Solid-State Circuits, IEEE Journal of*, vol.44, no.11, pp.3174,3192, Nov. 2009.

[2] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, K. Bernstein, “Scaling, power, and the future of CMOS,” *Electron Devices Meeting*, 2005. IEDM Technical Digest. *IEEE* International, vol., no., pp.7 pp.,15, 5-5 Dec. 2005.

[3] L. Chang, D.M. Fried, J. Hergenrother, J.W. Sleight, R.H. Dennard, R.K. Montoye, L. Sekaric, S.J. McNab, AW. Topol, C.D. Adams, K.W. Guarini, W. Haensch, “Stable SRAM cell design for the 32 nm node and beyond,” *VLSI Technol., Dig. Tech. Papers*, Jun. *2005*, pp.128,129.

[4] P. Zuber, M. Miranda, M. Bardon, S. Cosemans, P. Roussel, P. Dobrovolny, T. Chiarella, N. Horiguchi, A. Mercha, T.-Y. Hoffmann, D. Verkest S. Biesemans, “Variability and technology aware SRAM Product yield maximization,” *VLSI Technology (VLSIT), 2011 Symposium on*, vol., no., pp.222,223, 14-16 June 2011.

[5] S.O. Toh, Z. Guo, Liu, T.-J.King, B. Nikolic, “Characterization of Dynamic SRAM Stability in 45 nm CMOS,” *Solid-State Circuits, IEEE Journal of*, vol.46, no.11, pp.2702,2712, Nov. 2011.

[6] K.R. Lakshmikumar, R.A. Hadaway, M.A. Copeland, “Characterisation and modeling of mismatch in MOS transistors for precision analog design,” *Solid-State Circuits, IEEE Journal of*, vol.21, no.6, pp.1057,1066, Dec 1986.

[7] B. Razavi, “CMOS technology characterization for analog and RF design,” Solid-State Circuits, *IEEE* Journal of, vol. 34, no.3, pp.268-276, Mar. 1999.

[8] P.G. Drennan, C.C. Mcandrew, “Understanding MOSFET mismatch for analog design,” *Solid-State Circuits, IEEE Journal of*, vol.38, no.3, pp.450,456, Mar 2003.

[9] P.R. Kinget, “Device mismatch and tradeoffs in the design of analog circuits,” *Solid-State Circuits, IEEE Journal of*, vol.40, no.6, pp.1212,1224, June 2005.

[10] A. Asenov, “Simulation of statistical variability in nano MOSFETs,” *IEEE Symp. VLSI Technol., Dig. Tech. Papers*, Jun. 2007, pp. 86–87.

[11] K. J. Kuhn, “Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS”*, IEDM Tech. Dig*., pp.471 -474 2007.

[12] K. J. Kuhn, M. D. Giles, D. Becher, P. Koler, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai, “Process technology variation,” *IEEE Trans. Electron Devices*, vol. 58, no. 8, pp. 2197–2208, Aug. 2011.

[13] K. Takeuchi, A. Nishida, T. Hiramoto, “Random Fluctuations in Scaled MOS Devices,” *Simulation of Semiconductor Processes and Devices, 2009. SISPAD '09. International Conference on* , vol., no., pp.1,7, 9-11 Sept. 2009.

[14] T. Mizuno, J. Okamura, and A. Toriumi, “Experimental study of threshold voltage fluctuation due to the statistical variation of channel dopant number in MOSFETs,” *IEEE Trans. Electron Devices*, vol. 41, no. 11, pp. 2216–2221, Nov. 1994.

[15] C-H. Lin, R. Kambhampati, R.J. Miller, T. B. Hook, A. Bryant, W. Haensch, P. Oldiges, I. Lauer, T. Yamashita, V. Basker, T. Standaert, K Rim, E. Leobandung, H. Bu, M. Khare, “Channel doping impact on FinFETs for 22nm and beyond,” *VLSI Technology (VLSIT), 2012 Symposium on* , vol., no., pp.15,16, 12-14 June 2012.

[16] X. Zhang, J. Li, M. Grubbs, M. Deal, B. Magyari-Kope, B.M. Clemens, Y. Nishi, “Physical model of the impact of metal grain work function variability on emerging dual metal gate MOSFETs and its implication for SRAM reliability,” *in IEDM Tech. Dig.*, Dec. 2009, pp.1-4.

[17] D. Reid, C. Millar, S. Roy, A. Asenov, “Understanding LER-Induced MOSFET V_T Variability—Part I: Three-Dimensional Simulation of Large Statistical Samples,” *Electron Devices, IEEE Transactions on*, vol.57, no.11, pp.2801,2807, Nov. 2010.

[18] X. Sun, V. Moroz , N. Damrongplasit , C. Shin and T.-J. King Liu, “Variation Study of the Planar Ground-Plane Bulk MOSFET, SOI FinFET, and Trigate Bulk MOSFET Designs,” *Electron Devices, IEEE Transactions on*, vol.58, no.10, pp.3294, 3299, Oct. 2011.

[19] T. Tanaka, T. Usuki, T. Futatsugi, Y. Momiyama, and T.Sugii, “Vth fluctuation induced by statistical variation of pocket dopant profile”, in *IEDM Tech. Dig*., 2000, pp. 271-274.

[20] M. Miyamura, T. Nagumo, K. Takeuchi, K. Takeda, and M. Hane, “Effects of drain bias on threshold voltage fluctuation and its impact on circuit characteristics,” in *IEDM Tech. Dig*., Dec. 2008, pp. 447–450.

[21] T. Mizutani, A. Kumar, T. Tsunomura, A. Nishida, K. Takeuchi, S. Inaba, S. Kamohara, K. Terada, T. Hiramoto, “Statistic characteristics of ‘current-onset voltage’ in scaled MOSFETs analyzed by 8k DMA TEG,” *Silicon Nanoelectronics Workshop (SNW)*, 2010 , vol., no., pp.1,2, 13-14 June 2010.

[22] A. Cathignol, S. Bordez, A. Cros, K. Rochereau, G. Ghibaudo, “Abnormally high local electrical fluctuations in heavily pocket-implanted bulk long MOSFET,” *Solid-State Electronics*, Volume 53, Issue 2, February 2009, Pages 127-133.

[23] T. Hook, J. Johnson, J.-P. Han, A. Pond, T. Shimizu, and G. Tsutsui, “Channel length and threshold voltage dependence of transistor mismatch in a 32 nm HKMG technology,” *IEEE Trans. Electron Devices*, vol. 57, no. 10, pp. 2440–2447, Oct. 2010.

[24] C. Mezzomo, A. Bajolet, A. Cathignol, E. Josse, and G. Ghibaudo, “Modeling local electrical fluctuations in 45 nm heavily pocket-implanted bulk MOSFET,” *Solid State Electron*., vol. 54, no. 11, pp. 1359–1366, Nov. 2010.

[25] N. Damrongplasit, L. Zamudio, S. Balasubramanian, “Threshold voltage and DIBL variability modeling for SRAM and analog MOSFETs,” *VLSI Technology (VLSIT), 2012 Symposium on*, vol., no., pp.187, 188, 12-14 June 2012.

[26] M.J.M. Pelgrom, Aad C J Duinmaijer, AP.G. Welbers, “Matching properties of MOS transistors,” *Solid-State Circuits, IEEE Journal of*, vol.24, no.5, pp.1433,1439, Oct 1989.

[27] K. Takeuchi, T. Fukai, T. Tsunomura, A.T. Putra, A. Nishida, S. Kamohara, T. Hiramoto, “Understanding Random Threshold Voltage Fluctuation by Comparing Multiple Fabs and Technologies,” *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, vol., no., pp.467,470, 10-12 Dec. 2007.

Chapter 4

Variability Characterization in Fully-Depleted Silicon-On-Insulator (FD-SOI) Transistors

4.1 Introduction

In order to satisfy Moore's Law, transistors are made smaller in each successive technology node so that more of them can be put onto a chip [1-3]. However, such aggressive scaling can also have an adverse effect on the electrostatic integrity of a transistor, causing large off-state current and worsening short-channel effects. To mitigate such undesirable effects, planar bulk-silicon transistors often employ techniques such as the use of retrograde/halo doping, shallow source/drain junctions, and high-k metal gate stacks [4-7]. One of the root causes of poor electrostatic control is relatively weak capacitive gate coupling to the electric potential in the silicon body region that is further from the gate-oxide interface. The Si region which is furthest away from the gate can serve as a major leakage path [8-9]. To tackle this challenge head on, one can think of removing all paths far away from the gate. In fact, this is precisely the idea behind the thin-body (fully depleted) MOSFET [10-14]. If the thickness of the silicon body is made much thinner than the gate length short-channel effects are dramatically reduced. The two most common implementations of a thin-body MOSFET today are the vertical FinFET or planar FDSOI (Fully-Depleted Silicon-On-Insulator) MOSFET [15-16].

The FinFET is a double-gate MOSFET structure which is more scalable compared to the FDSOI MOSFET due to superior gate control. However, it requires a high aspect ratio Si fin geometry, which presents a major challenge from a fabrication standpoint. Additionally, since the drive strength of a FinFET is adjusted by changing the number of fins, circuit designers must cope with discrete adjustments in drive current for FinFETs [17]. On the other hand, the FD-SOI MOSFET structure, which also uses a thin body similar to the FinFET, offers improved electrostatic control over the planar bulk MOSFET without adding significant fabrication challenges or imposing new restrictions on circuit design. Instead of a bulk Si wafer, the starting substrate is a Silicon-On-Insulator (SOI) wafer [18]. The device fabrication process steps are very similar or less complicated compared to those of a standard planar bulk Si device fabrication process. From a circuit designer standpoint, the FDSOI design kit is also similar to that for bulk Si technology: device widths can be adjusted to tune transistor drive strength, and back-biasing can be used to dynamically adjust transistor threshold voltage [19-20].

Given that FDSOI technology is a promising candidate to replace planar bulk Si technology, variability analysis of FDSOI MOSFETs is necessary. This can be achieved by implementing a device characterization array and padded-out SRAM cells in a test chip.

4.2 Device Characterization Array

To capture and understand the impact of different variability sources on device performance, transistors of different sizes and layout geometries are included in the device characterization array. Ring oscillators, capacitance test structures, and resistance test structures can also be included. Using built-in circuitry on the chip, each individual device can be electrically accessed and characterized through the input/output (I/O) pads. In general, the variability test structures can be classified as 1) Random or 2) Systematic variability test structures. Both NMOS and PMOS transistors, of different V_T values, are included. Test chips were fabricated by STMicroelectronics using a 28nm high-k/metal-gate (HKMG) process technology, on both bulk-Si and SOI substrates to allow for a direct comparison of planar bulk *vs*. FDSOI technologies. The layout of the device characterization block is shown in Fig. 4.1

.

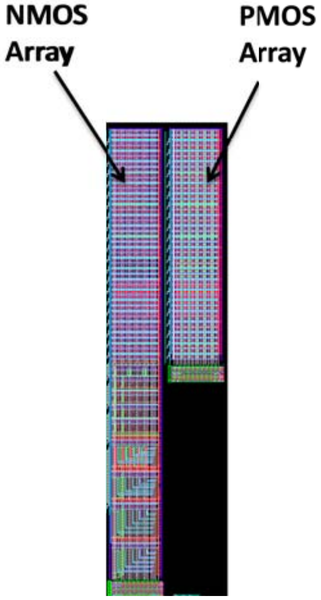


Fig. 4.1 Layout view of the device characterization block consisting of NMOS and PMOS transistor arrays.

4.2.1 Random Variability Test Structures

Random variability sources such as random dopant fluctuations (RDF), gate work function variation (WFV), and line-edge roughness (LER) can contribute to variations in V_T,

I_{OFF}, and I_{ON} between devices with identical layouts. In order to isolate the impact of random variability from that of systematic variability, transistor pairs (i.e. mismatch test structures) are often used. These test transistors are identically drawn structures and they are placed in close proximity to one another on the chip. If there is a systematic source of variability, its impact would be the same for both devices. As a result, when the difference (as opposed to the absolute value) of the performance parameter between the two transistors in a pair is analyzed, the impact due to systematic variability is canceled out, i.e. the difference is due entirely to random variability. To ensure that the transistors in a pair are identical in every possible aspect, it is important to make sure that the surrounding area is the same for both transistors. Fig. 4.2 shows a Device Under Test (DUT) surrounded by dummy active regions. The other corresponding DUT in the pair is also drawn in a similar manner. Such a layout will help to eliminate variability that might arise from layout-dependent proximity effects such as mechanical stress from Shallow Trench Isolation (STI) or near-by active devices [21-22].

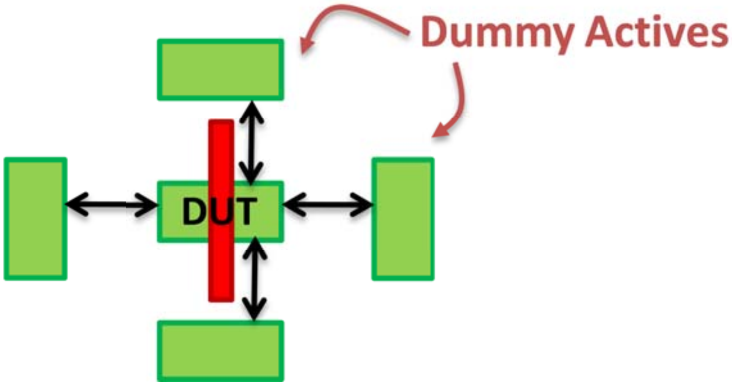
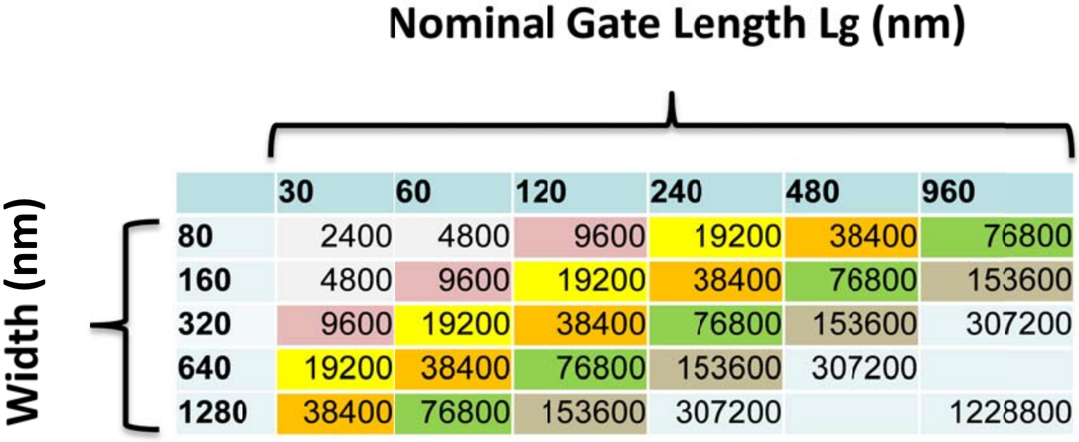


Fig. 4.2 A layout showing Device Under Test (DUT) surrounded by dummy active regions with equal distance to elimate any layout-dependent proximity effects.

It has been theoretically derived and experimentally validated that variability in MOSFET threshold voltage increases as the transistor channel dimensions are made smaller. Specifically, σV_T is proportional to $1/\sqrt{W \cdot L}$ [23-25]. Thus, devices with different channel area values (i.e. W x L), ranging from large to small, are included in the array to assess the significance of this trend. The list of transistor channel dimensions is summarized in Fig. 4.3. To further examine the sensitivity of variability sources to various device design parameters (e.g. channel width and gate length), different combinations of W and L corresponding to a fixed channel area are also included; devices which have the same channel area are coded with the same color in Fig 4.3. This allows one to decouple different variation sources affecting the transistor threshold voltage. For example, the impact of gate LER on σV_T might be more sensitive to gate length scaling as compared to channel width scaling. Thus, different Pelgrom plots can be generated to compare the effect of scaling the channel length only or the channel width only, while keeping W x L constant.

For System-On-Chip (SOC) products, multiple values of V_T must be available to the designers [26-27]. Therefore, it is also important to investigate how variability will affect devices of different nominal V_T values. To this end, three different V_T levels (Low V_T, Regular $V_{T,}$ and

High V_T) are included for each value of W x L. V_T tuning is achieved through a combination of gate length trimming and backplane doping underneath the Buried Oxide layer (BOX).



Nominal Gate Length Lg (nm)

Width (nm)

	30	60	120	240	480	960
80	2400	4800	9600	19200	38400	76800
160	4800	9600	19200	38400	76800	153600
320	9600	19200	38400	76800	153600	307200
640	19200	38400	76800	153600	307200	
1280	38400	76800	153600	307200		1228800

Fig. 4.3 Summary of MOSFET channel dimensions included in the device characterization array.

4.2.2 Systematic Variability Test Structures

In addition to device structures used to study random variability, several device structures are included to assist with the study of systematic variability associated with layout proximity effects, including mechanical stress induced by STI, Length of Diffusion (LOD), well doping proximity, and segmented channel design. The design and layout of these structures are summarized in the following sections.

4.2.2.1 Shallow Trench Isolation (STI) Effect

Mechanical stress induced by STI can affect carrier mobility and thereby affect transistor on-state drive current [28-29]. To quantify the impact of STI-induced stress from different directions, dummy active regions are drawn at different distances (λ) away from the device under test. For example, to examine the stress induced along the channel direction of a transistor, the two dummy active rectangles located on the sides of the DUT are drawn at distances of λ, 2 λ, 3 λ, and 4 λ away from DUT, as shown in Fig. 4.4. Similarly, the effect of STI-induced stress across the channel (along the width direction) can also be captured by placing the dummy active regions at the top and bottom of a DUT at different distances as depicted in Fig. 4.5.

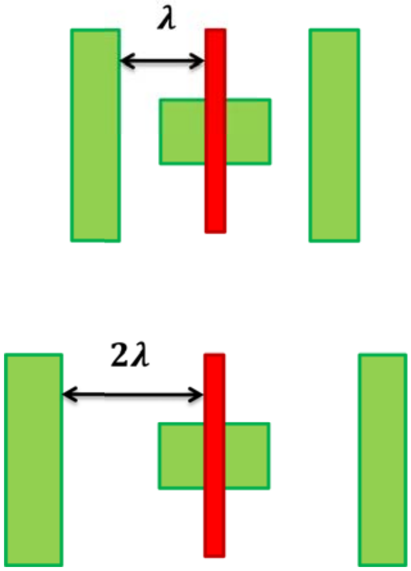


Fig. 4.4 Test structures to monitor the effect of STI-induced stress along the channel direction (lateral).

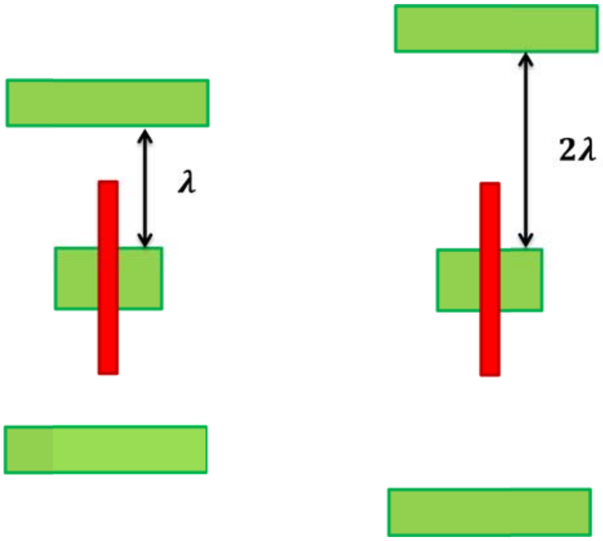


Fig. 4.5 Test structures to monitor the effect of STI-induced stress across the channel direction (vertical).

4.2.2.2 Gate Effect

Design Rule Check (DRC) often requires that a minimum density of the gate/poly layer is achieved in order to ensure an acceptable yield during the chemical-mechanical polishing (CMP) step [30]. However, these dummy poly structures or gate electrodes of neighboring transistors can influence the stress within the channel region of the DUT. In addition to stress effects, optical proximity effects during the photolithographic exposure process can cause the adjacent gate electrodes to have an impact on the patterned shape of the gate electrode [31]. The test structures used to investigate the impact of this neighboring gate/poly feature effect are shown in Fig. 4.6, where one device has a dummy poly feature on each side of the DUT and the other device only has a dummy poly feature on one side.

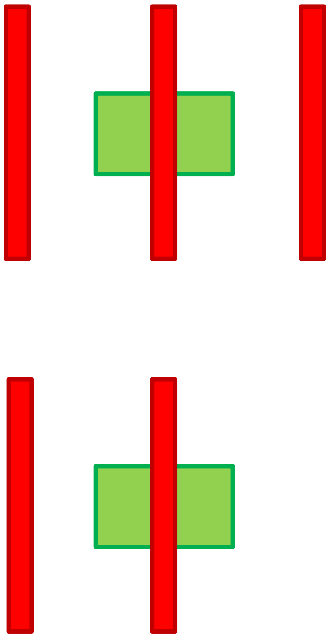


Fig. 4.6 Test structure used to study systematic variability induced by neighboring gate-level features.

4.2.2.3 Length of Diffusion

The stress profile within the channel region of the DUT also depends on the length of the diffusion (LOD) or source/drain regions of the transistor [32]. To study the impact of LOD on device performance, transistors with same gate length L_g and channel width W were drawn with different diffusion lengths at λ, 3λ, 4λ, and 5λ for both the source and the drain sides, as shown in Fig. 4.7.

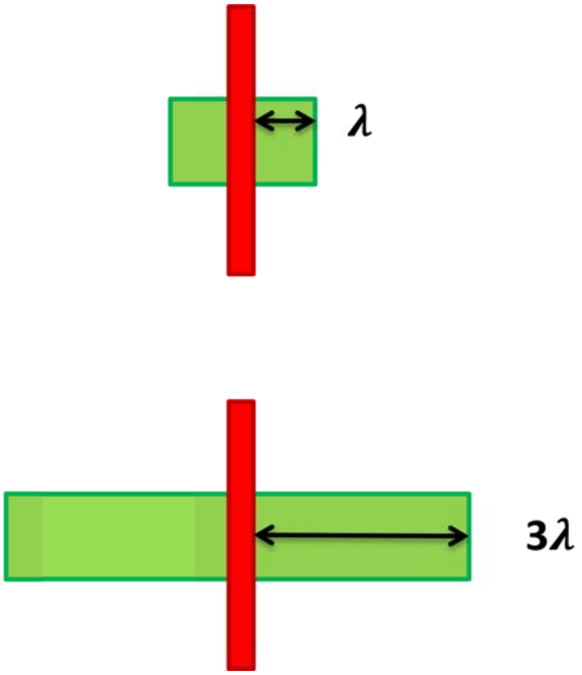


Fig. 4.7 Test structures used to study the impact of length of diffusion on transistor performance.

4.2.2.4 Asymmetric Source/Drain Diffusion Length

In addition to devices having equal source/drain diffusion lengths, the LOD of a DUT can also be asymmetric (e.g. source-side LOD is longer than drain-side LOD), as illustrated in Fig. 4.8. This test structure can be used to decouple the impacts of source-side LOD vs. drain-side LOD, permitting a close examination of parameters which are sensitive to S/D asymmetry such as $V_{T\ SAT}$ and source-injection velocity.

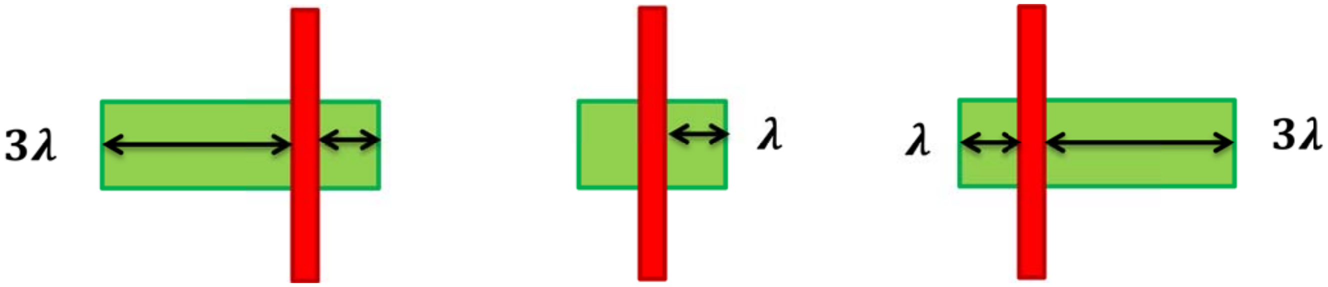


Fig. 4.8 Test structures with asymmetric source/drain diffusion lengths.

4.2.2.5 Shared Source/Drain Mismatch Pair

Two test transistors are drawn such that a diffusion region is shared between them, as depicted in Fig. 4.9. This particular layout is common when transistors are connected in a stack or cascade configuration. This structure is quite useful for random variability study since the two structures are as close as possible to one another. It can also be used to monitor systematic variability such as the difference in channel stress profiles within DUT#1 and DUT #2.

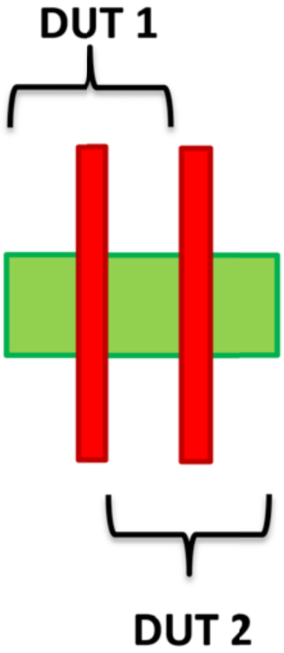


Fig. 4.9 Mismatch pair test structures with shared source/drain region.

4.2.2.6 Well Proximity Effect

In a planar bulk CMOS technology, NMOS and PMOS transistors are placed inside P-well and N-well regions, respectively [22, 33]. In a FD-SOI CMOS technology, the doping type underneath the isolating buried oxide (BOX) layer can be adjusted to achieve the desired V_T specification, for both NMOS and PMOS transistors. Due to lateral straggle of implanted dopant atoms, the doping concentration within the well region of the DUT can be affected if it is situated close to the boundary between the N-well and the P-well. To investigate this effect, test structures shown in Fig. 4.9 and 4.10 are used, wherein the DUTs are placed at different distances away from the boundary of the well doping, laterally as well as vertically. In order to isolate the well proximity effect from a particular direction, devices are placed at least 3 um away from that particular well boundary. For example, to example the effect of the side N-well on NMOS transistors residing within a P-well, DUTs are placed at distances of λ, …, 6λ away from the side N-well/P-well boundary, while ensuring that all of the DUTs are at least 3 um away from the bottom N-well.

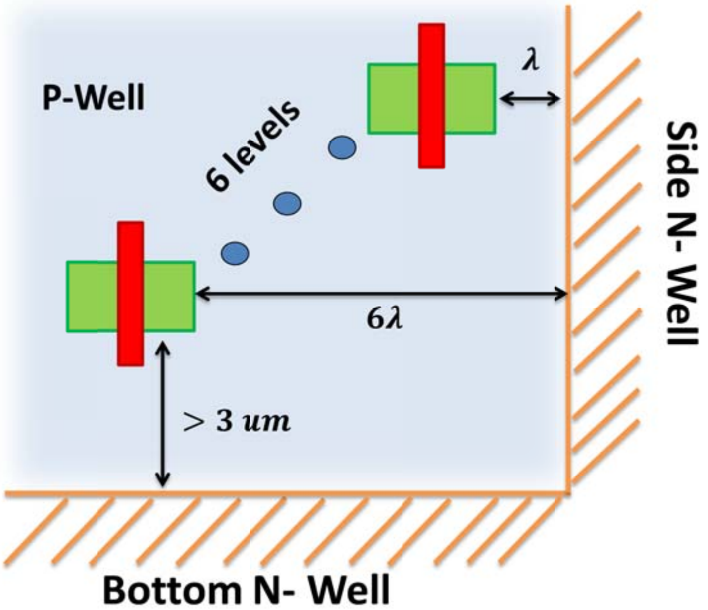


Fig. 4.9 Test structure used to study the impact of proximity to the side N-well. Note that DUT are placed at least 3 um away from the bottom N-well boundary to ensure that this bottom well does not affect the DUTs.

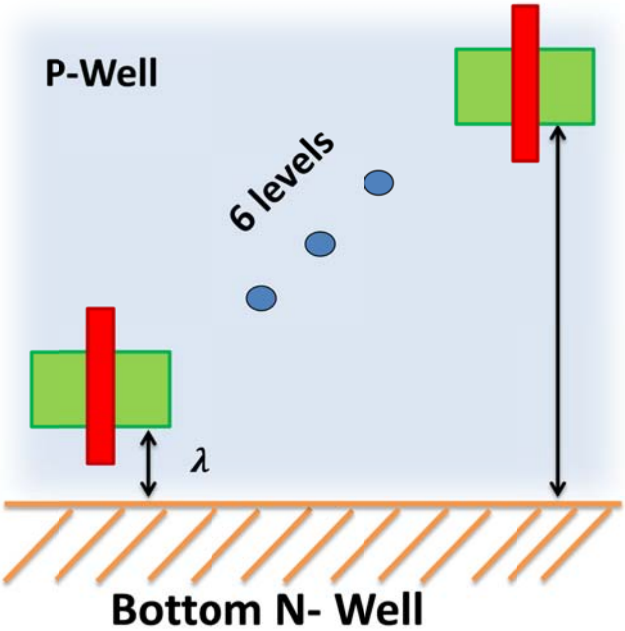


Fig. 4.10 Test structure to study the impact of proximity to the bottom N-well.

4.2.2.7 Segmented Channel Transistor

Instead of a transistor having a continuous width, the channel region can be segmented into multiple stripes of equal width as shown in Fig. 4.11. From an electrostatic control standpoint, a segmented channel transistor can offer improved short-channel effect due to the slight wraparound of the gate over the channel (i.e. a common occurrence in the STI process due to HF over etch of the STI oxide) and the gate fringing electric field coupling to the channel region through the STI [34-36], if the stripe width is comparable to the channel length. Thus, even though the segmented channel design takes up more layout area as compared to a conventional channel design, the improvement in device performance (subthreshold swing, DIBL, I_{ON}, I_{OFF}) can provide a net benefit when normalized to the same layout area. Additionally, larger and more uniform mechanical stress can be induced within narrow channel segments. In order to observe the greatest benefit of the segmented channel design, the minimum drawn device width is used for each channel segment.

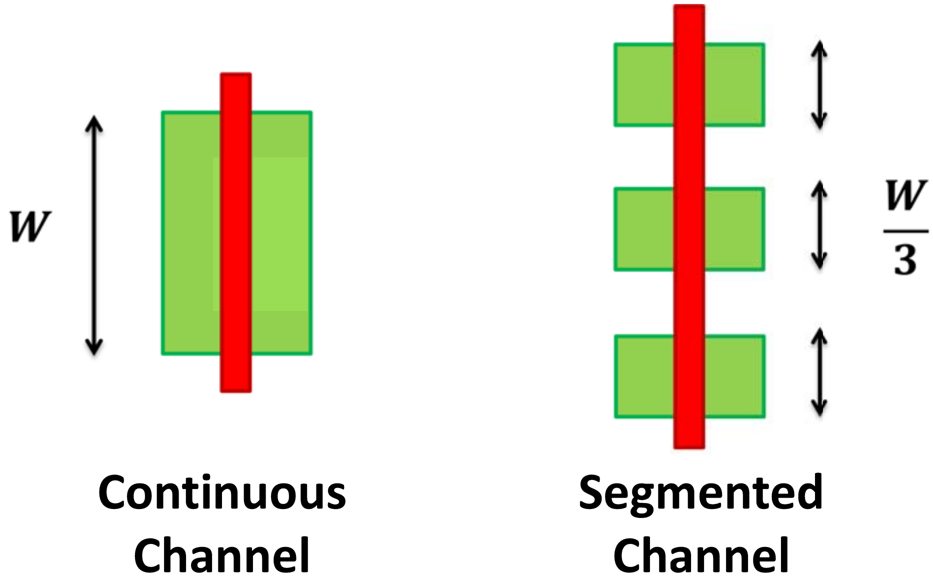


Fig. 4.11 Test structures comparing continuous channel *vs*. segmented channel designs.

4.2.3 Test Circuitry

Device characterization arrays or "blocks" allow for a large number of test devices of different designs to be included on a single die. Usually, these test devices are not directly probed due to the limited die area. Moreover, without an industrial grade auto prober, directly probing each device in order to collect statistical data on device variability can be a very time consuming task. Therefore, electrical access to individual devices in the array is made through the Input/Output (I/O) pads of the test chip. Since the number of I/O pads is limited and must be sufficiently allocated for all of the signals in different test blocks, a decoder circuit is used to share some common digital signals such as scan-in (S_{IN}), scan-out (S_{OUT}), and scan –clock

(S_{CLK}). The decoder is controlled by a 2-bit signal denoted as ‘Sel<0:1>’. In the device characterization block, there are 4 main characterization modules: NMOS array, PMOS array, C-V structures, and Ring oscillator. A digital select signal is used to select one of these 4 modules, and only one of the modules is active at any given time. Additionally, an enable signal (EN) for device characterization is also included. If the EN signal is off, the device characterization block is turned off, preventing it from interfering while the SRAM or microprocessor blocks are being tested. The floor plan of the device characterization array is shown in Fig. 4.12. There are 6 columns each in the PMOS and NMOS modules: Mismatch pair transistors of RVT, LVT, and HVT flavor each take up two columns. The digital signals and the selection circuitry are summarized in Fig. 4.13.

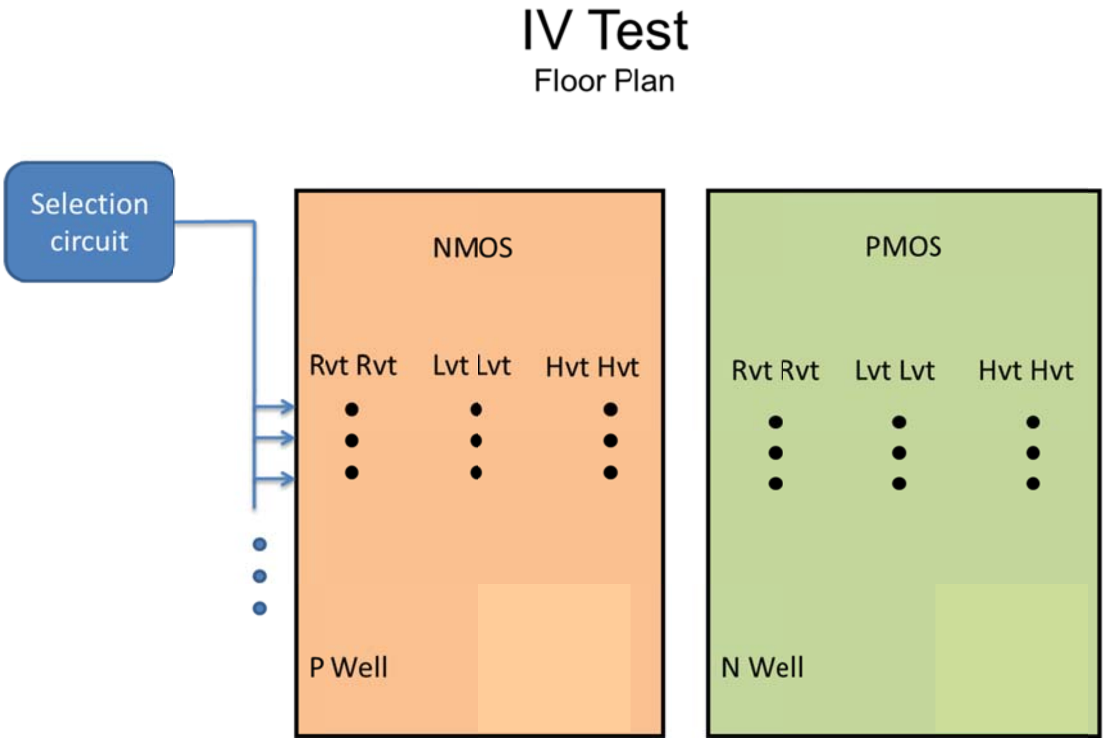


Fig. 4.12 Floor plan for the device characterization array consisting of selection circuitry, NMOS array, and PMOS array. In each array, mismatch transistor pairs for each V_T flavor (RVT, LVT, HVT) are included.

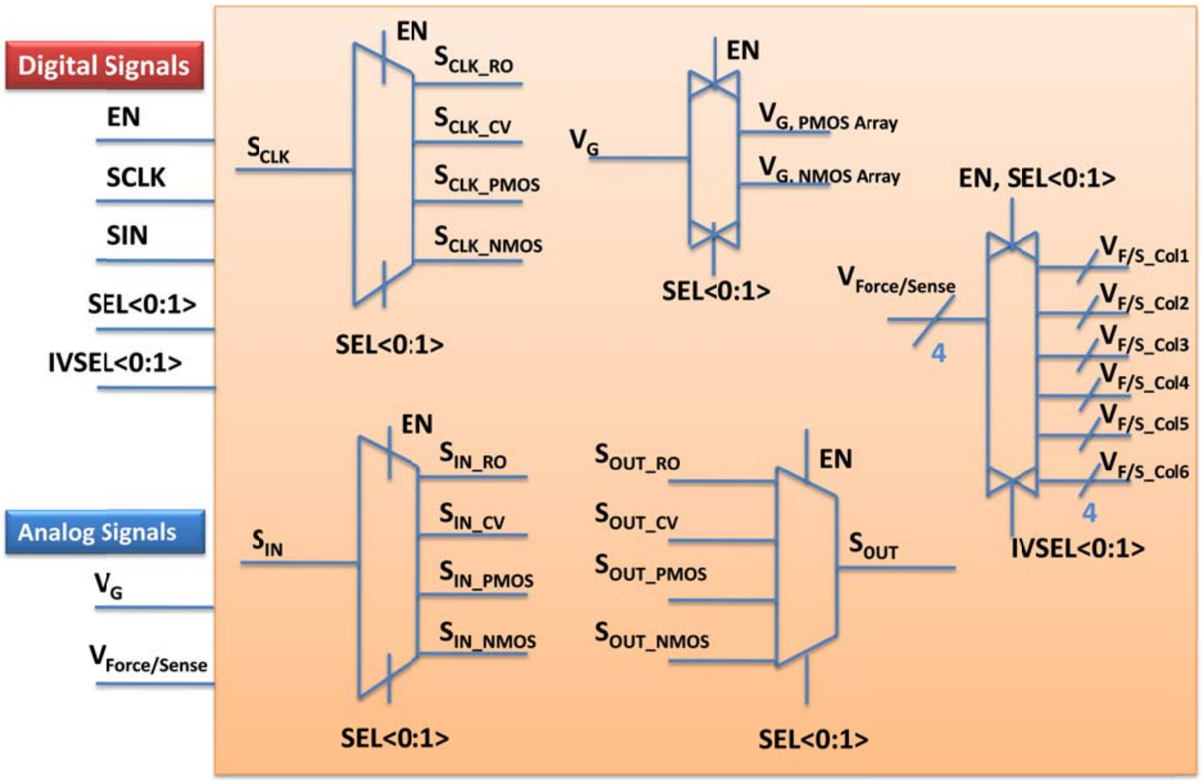


Fig. 4.13 Top-level schematic showing selection circuitry and analog signals for the device characterization array.

Devices within an array are accessed in a serial manner, in a row-wise fashion using a scan-chain circuit. A simple scan-chain circuit comprised of D flip-flops (D-FF) chained together is shown in Fig. 4.14. The output at each of the FF (i.e. S_1 and S_2) can be used to activate/deactivate the row under test. To prevent a race condition, from a layout standpoint it is a good practice to send the scan-IN signal at one end of the chain and to send the scan-CLK signal at the opposite end. A particular column can be selected through a column multiplexer that has a 3-bit control signal called 'IV_Sel<0:2>', allowing one out of the six columns to be selected at a given time. The source and drain of devices in the same column share electrical lines. With a combination of row select (through scan-chain clocking) and column select (through multiplexer), each individual transistor inside the array can be accessed, allowing full control over the biasing of the source, drain, and gate terminals through the I/O pads.

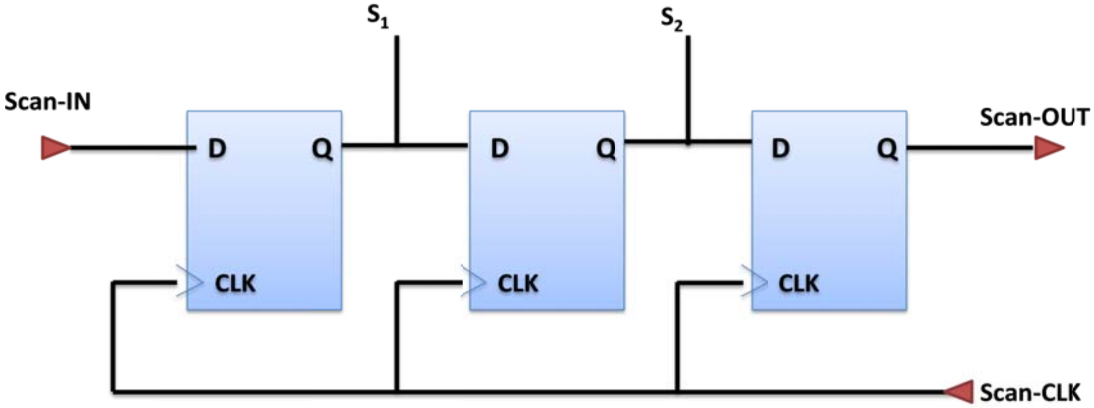


Fig. 4.14 Circuit schematic of a basic scan-chain. The schematic reflects the design layout where scan-IN and scan-CLK are fed into opposite ends of the chain to prevent a race condition.

Analog signal lines used to apply voltage to and measure current from the terminals of the DUT are routed through a series of pass-gates. Due to parasitic resistance along the wire trace from the probe pad to the DUT, the voltage applied to the device's terminals is smaller than the voltage applied at the pads. This voltage difference can be significant if resistance along the wire is large. To circumvent this problem, the Kelvin measurement technique utilizing separate sets of Force and Sense lines is implemented for the source and drain terminals of the DUT. Fig. 4.15 shows the basic Kelvin measurement configuration for the source and drain terminals. The current is passed through the Force lines and the voltage drop across the DUT is sensed across the Sense lines. Ideally the sense lines should be as close as possible to the DUT. In this case, the preferred sense lines would be the two inner lines which are common to the DUTs in the same column. The access transistors which are used to connect the force lines to the device in a particular row must be large enough to support the current level of the DUT, but small enough such that they will not have large off-state leakage current. Due to the high impedance associated with the sense line (similar to the impedance of a voltage meter), very low parasitic current can flow through it. The negative feedback mechanism inside the Source Measurement Unit (SMU) forces sufficient current through the device until the target voltage is reached across the sense lines.

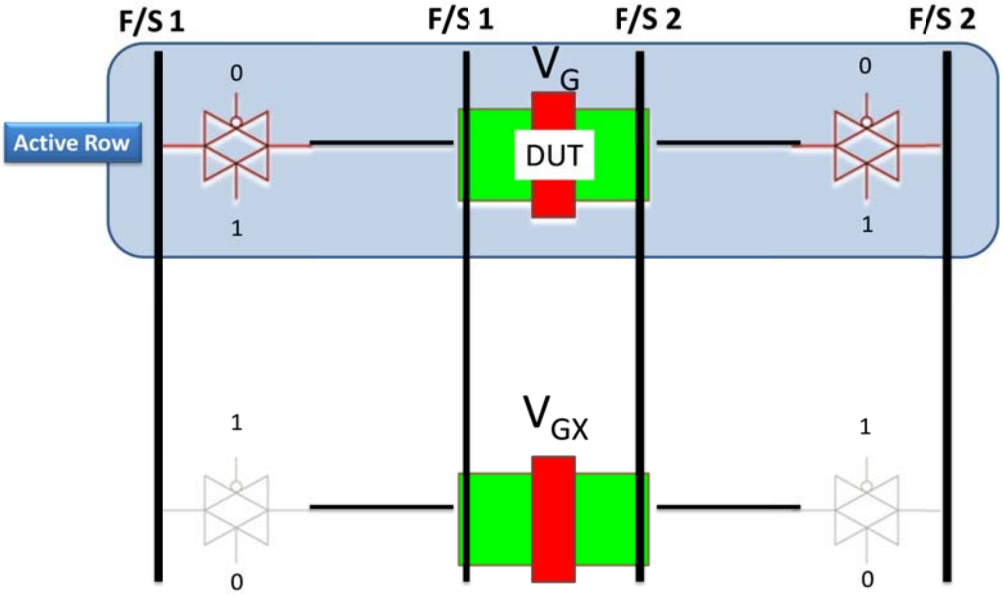


Fig. 4.15 Circuit schematic showing access transistors for row selection and Kelvin Force/Sense configuration.

Leakage control is also important in this type of device characterization array. Since only one active device is to be characterized at a time, the leakage current from the other devices in the same column should be as small as possible. To this end, a separate gate bias voltage (V_{GX}) is applied to turn off the transistors which are not being tested. This V_{GX} can be set to be negative or positive to make the NMOS or PMOS devices more strongly off, respectively. It is important to note that the maximum V_{GX} value allowed will be limited by Gate-Induced Drain Leakage (GIDL), which increases with negative (positive) gate voltage in NMOS (PMOS) transistors. Fig. 4.16 shows the schematic of the leakage control circuit. Note that it is very critical to appropriately choose the size of the NMOS transistor used to pull the gate voltage of an idle device down to a negative voltage V_{GX}. This can be problematic because when EN is 1, the NMOS device that is used to pass the negative gate bias V_{GX} should be off since its gate voltage V_G = 0V; but because its source voltage is negative ($V_S = V_{GX}$), V_{GS} is positive and hence the transistor is weakly on. To compound this problem, the body terminal of the transistor is usually biased at GND for NMOS and V_{DD} for PMOS, therefore there exists a small forward body bias which lowers the V_T of this NMOS transistor even more. Ultimately, this can affect the voltage at the gate terminal of the DUT, since the pass gate is trying to set the node voltage to V_G but the pull-down NMOS transistor is trying to set it to V_{GX} instead. Therefore, the effective gate voltage that the DUT sees will be smaller than V_G and it will not increase linearly when one tries to linearly sweep the gate voltage.

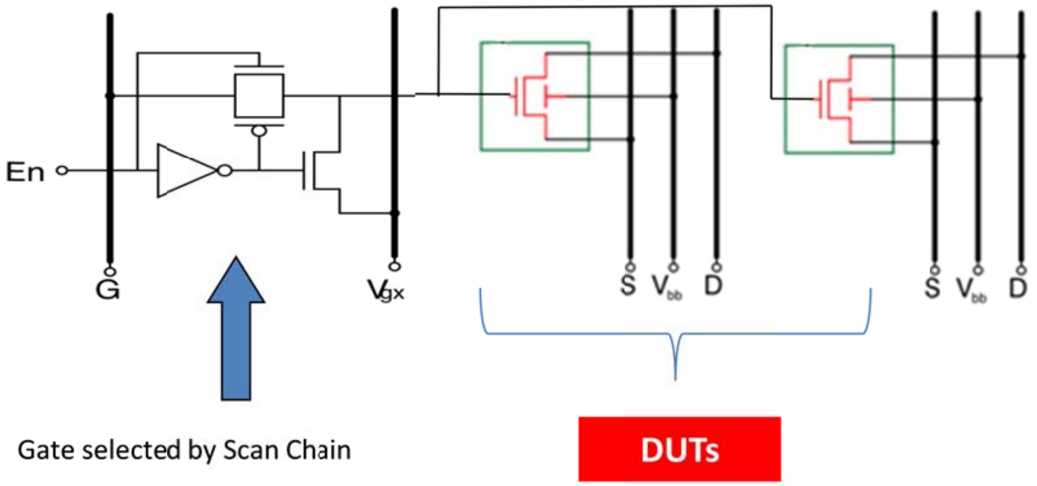


Fig. 4.16 Row selection circuit with V_{GX} biasing for minimizing off-state leakage for an NMOS array.

4.2.4 Test Setup and Measurement

The fabricated chip was packaged in a pin grid array (PGA) mountable to a printed circuit board (PCB) using a standard chip socket. The PCB was designed in-house and it contains various test pins, decoupling capacitors, voltage level shifter, and connectors for digital and analog signals. Connectors for analog signals are coaxial for gate terminal biasing, while triaxial connectors are used to bias the source and drain terminals. (Triaxial connectors help to prevent leakage current through the insulator of the cable, which is necessary for measuring small levels of current flowing through the source and drain of a device.) A photograph of the test chip mounted on the PCB is shown in Fig. 4.17.

A Semiconductor Parameter Analyzer (Agilent B1500) is used to control the digital and analog signals. Specifically, digital signals used in the selection circuitry are inputted through the DB25 port of the Agilent B1500. The SPA outputs the signal according to the binary representation of the programmed decimal value. For this specific instrument, digital high is represented as a ‘0’ and digital low is represented as a ‘1’, which is contrary to the normal standard. A level shifter is needed on the PCB board to convert the digital output from the Agilent B1500 (5V) to the acceptable range used by the test chip (0.9V- 1V). An example of a scan-chain test with signals coming out of DB25 is shown in Fig. 4.18.

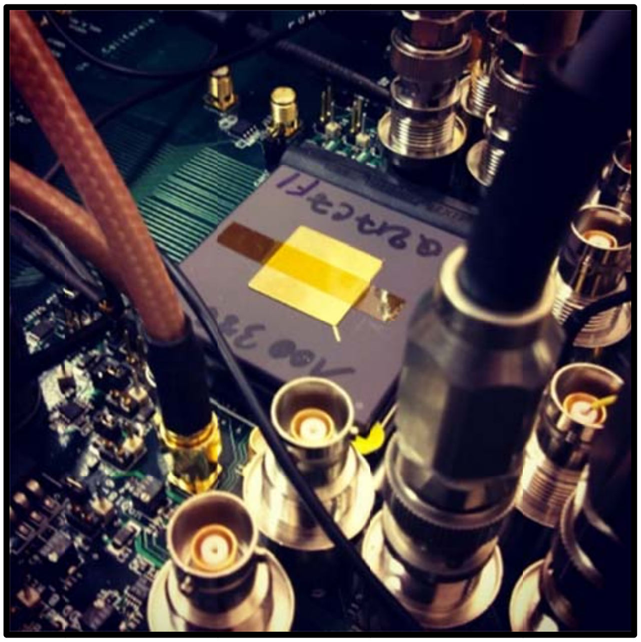


Fig. 4.17 Packaged test chip mounted on the custom-made PCB. Coaxial and Triaxial cables are used for delivering analog signals to the chip.

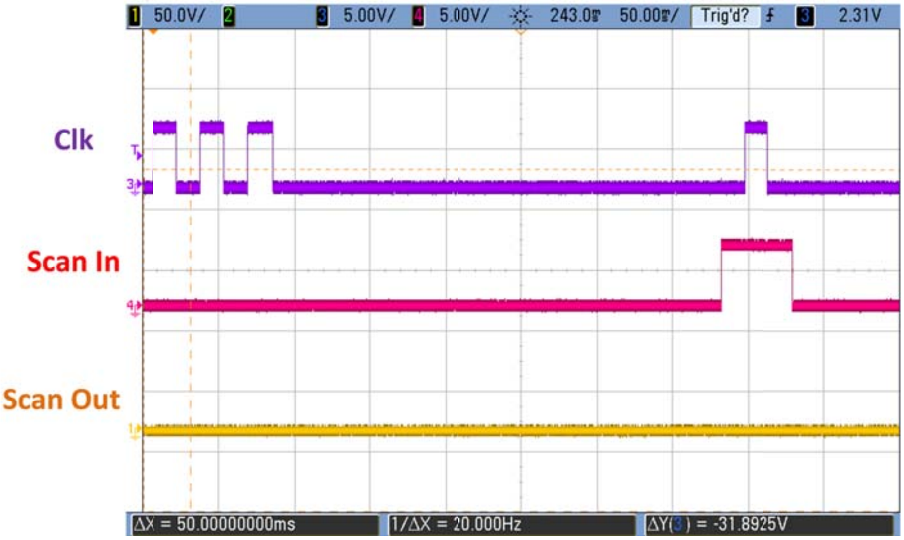


Fig. 4.18 Digital signals outputted from DB25 port used to control the scan-chain.

Once the device in the array is selected, its source, drain and gate terminals are multiplexed to the I/O pads which are now connected to the SPA. This allows one to perform any basic electrical measurements such as I_D-V_G and I_D-V_D. However, the I-V characteristic tends to suffer from a high leakage floor due to the off-state leakage currents flowing through the other transistors within the same column of the characterization array as well as the devices in the selection circuitry (pass gates, logic gates, and multiplexers).

As mentioned in the previous section, the built-in leakage control circuitry allows the gate terminal of the idle transistors to be set at V_{GX} in order to make them strongly off. In conjunction with this method, one can also try to calibrate out the leakage current. This is accomplished by first performing a parametric I-V sweep with none of the devices in the array selected (i.e. the scan-chain is filled with zeros). Let's call the resulting current $I_{Leakage}$. Then, one can perform the I-V measurement for the DUT to get I_{DUT}. To calibrate the leakage current out, the two current quantities are subtracted from one another, i.e. $I_{DUT,Calibrated} = I_{DUT} - I_{Leakage}$. The quality of the calibrated current depends on the current sensitivity level of the SPA which can be set for optimal operation.

4.3 SRAM Characterization Array

Static memory (SRAM) is a critical component of VLSI systems today. SRAM can provide the fastest random access time to stored data, and is used for lower-level caches (L1-L3) and registers [37]. In order to increase the size of the cache on a chip, it is desirable and economical to fit as many cells into an SRAM array as possible. However, as the memory cell area is scaled down with each new technology node, the read and write margins are degraded due to increasing variability in transistor characteristics. It is desirable to minimize the operating voltage V_{DD} of an SRAM cell in order to minimize power consumption. But a small mismatch in V_T values can significantly reduce cell stability, setting a lower limit for the minimum operating voltage $V_{DD,\ min}$ of the cell [37-39]. Therefore, variability analysis for SRAM is important for improving cell yield to reduce $V_{DD,\ min}$.

A widely used SRAM cell design is the six-transistor (6T) cell, consisting of one pair of NMOS pull-down transistors, one pair of PMOS pull-up transistors, and one pair of NMOS pass-gate transistors. Since the transistors in the SRAM cell are packed very close to one another to maximize storage density, the transistor pairs inside a 6T cell naturally form mismatch pairs which are ideal for studying random variability.

4.2.1 Padded-out 6T SRAM Macro

A SRAM cell characterization block consisting of 6T SRAM cells was designed for a 16nm FD-SOI CMOS technology developed by CEA-LETI. There are 14 SRAM macros in total. Each consists of a 128 kB array and peripheral circuitry. Different flavors of SRAM macros are included, with design variations in transistor sizes and threshold voltage values. The floor plan for the 14 different SRAM macros is shown in Fig. 4.19. For each macro, 21 memory cells inside the array are fully padded out. Fig. 4.20 illustrates the approximate location of the padded out cells with respect to the entire array, containing 10 horizontal cells, 10 vertical cells, and 1 center cell. The locations of the padded out cells are chosen so that systematic variation across the SRAM array can be effectively monitored. Additionally, these particular cell locations can minimize some variation gradients and isolate stress variation within the array.

In a normal 6T SRAM cell, the only accessible nodes are the Word Line (WL), Bit Lines (BL, BLB), voltage supply (V_{DD}), and ground (GND). This is acceptable if one just wants to perform basic SRAM operations such as read, write, and hold. However, in order to gauge SRAM cell stability by generating a butterfly plot, it is necessary to access the internal storage node in order to sweep the voltage on the node. To this end, padded out SRAM cells are employed such that every node of a transistor inside the cells can be directly accessed. Not only does this permit the butterfly plot to be generated, but also the standard transistor level characterization test such as current *vs.* voltage can be performed, allowing direct correlation between transistor performance parameters and SRAM cell metrics. To eliminate the effect of parasitic resistance along the wires, the Kelvin Force and Sense configuration is implemented for each of the terminal in the padded-out 6T SRAM cells. All of the wiring in the SRAM array is formed using only two metal layers.

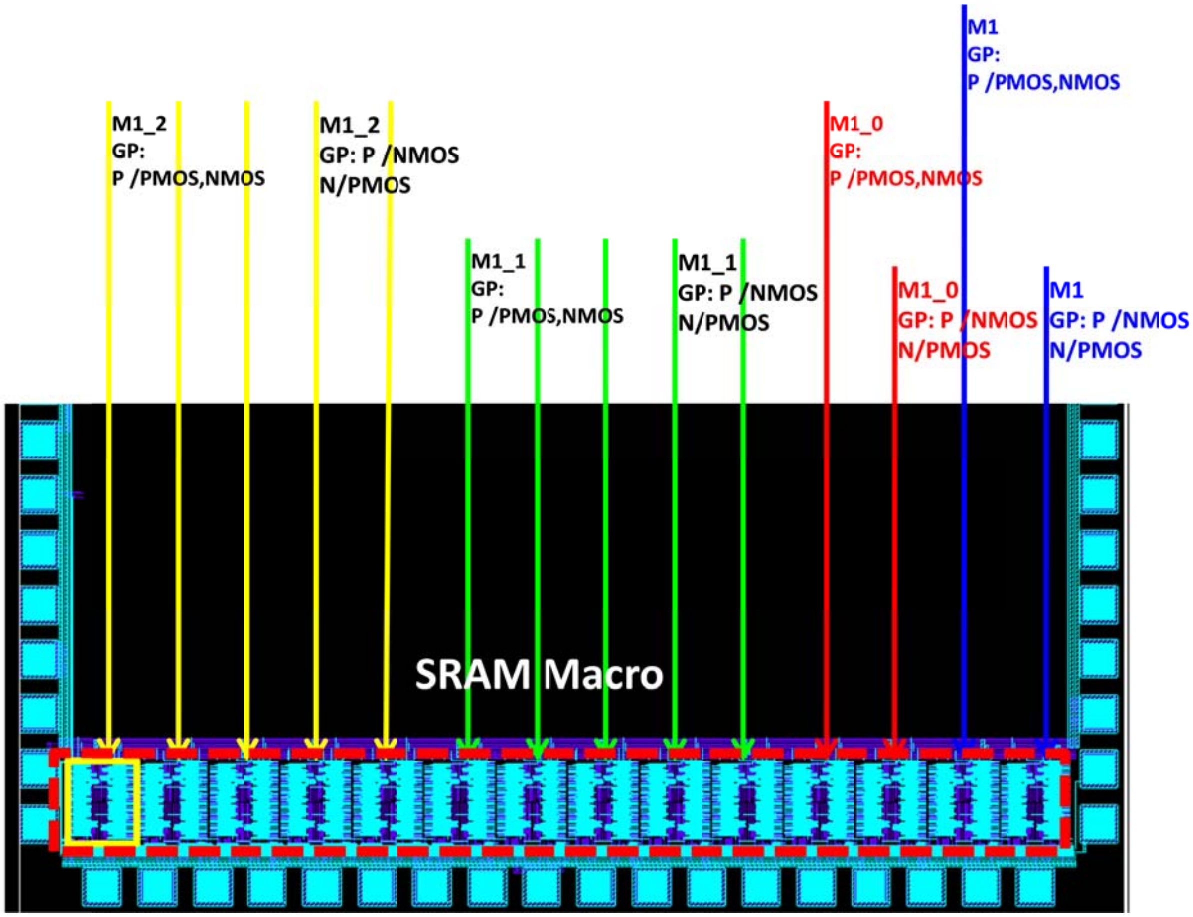


Fig. 4.19 Schematic layout of the SRAM macros designed for 16nm FD-SOI CMOS technology.

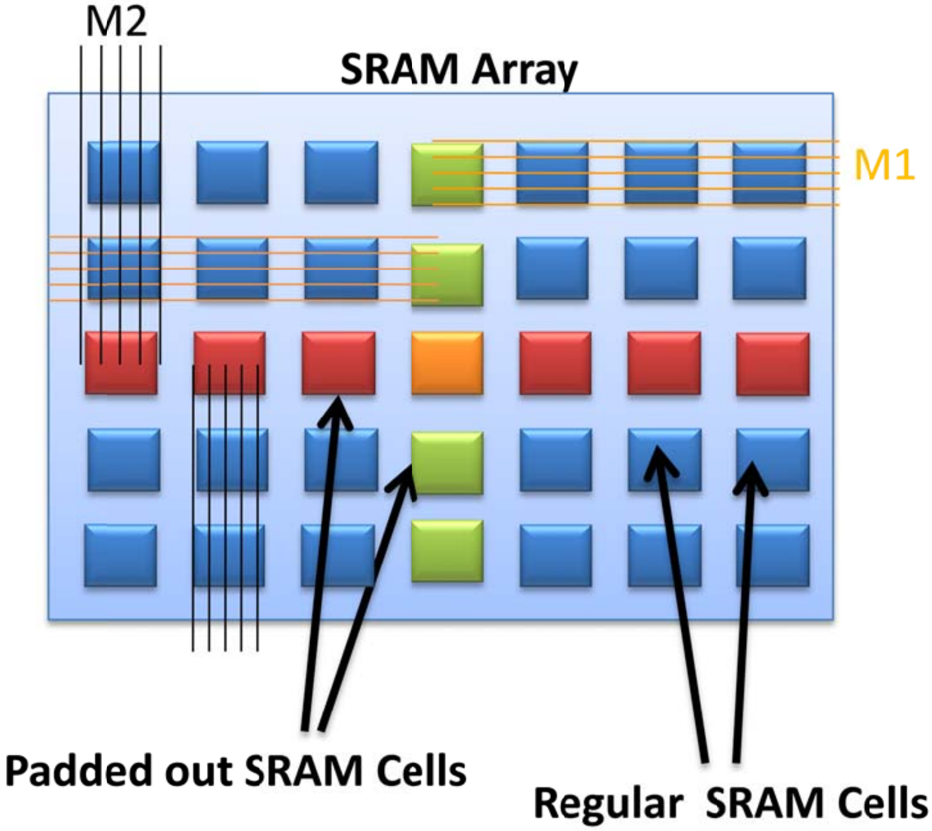


Fig. 4.20 Locations of padded-out SRAM cells inside the SRAM macro array.

4.2.2 Test Circuitry

Each one of the padded-out SRAM cells can be accessed serially through a scan-chain. Since there are 14 macros and each one contains 21 padded-out cells, the length of the scan chain is 14x21 = 294.

A switch box is the central circuitry in the SRAM macro. Its main purpose is to implement different modes for a cell that has been selected by the scan-chain. These operations include assigning different bias voltages to a particular node inside the 6T cell. This allows for I-V measurement of any one of the 6 transistors inside the cell. The digital control signal that is fed to the switch box is called switch box select 'swsel', and is implemented using a separate scan-chain where the output of each stage makes up the swsel signal. Once a particular set of binary sequences has been loaded into the swsel chain, logic gates are set such that the desired operation is performed. For example, if the binary sequence is meant to enable a I_D-V_G sweep of a PG transistor on the left half of the SRAM cell, the circuit will use this binary sequence to connect the source, drain, and gate terminals of the left PG transistor to the correct analog signals on the I/O pads. The adjacent macro array shares the switch box in order to minimize footprint.

In addition to regular I-V characterization, a voltage stress mode is also implemented on this test chip. NBTI/PBTI and random telegraph noise (RTN) can have a significant impact on device variability over time [40-42]. In order to characterize these effects, one would start out by measuring the I-V characteristic of an unstressed device. Then, a large voltage is applied to stress the device for a period of time. And immediately following the stress period, another I-V sweep is performed to examine the device characteristic just after stress. Once the measurement has been taken, the device can be placed back in stress mode for continued monitoring of stress-induced degradation. The short amount of time required to switch between I-V mode and stress mode is made possible through the switchbox logic. Fig. 4.21 shows a circuit schematic of the devices in stress mode. The bias conditions are chosen carefully such that a pair of transistors (NMOS and PMOS) can be stressed at the same time for each cell. For example, to stress PD1 and PU2, the logic circuit would pass V_{stress} (typically greater than 1V) to the gate of PD1 and 0V to the CL node. Effectively, PD1 will have 0V at both its source and drain, and V_{stress} at its gate. To piggy back on this biasing scheme, PU2 can also be put under stress at the same time as PD1. Since the gate of PU2 is the same as the CL node, it will also have a 0V on it. One terminal of the source /drain of PU2 is already connected to CH node, which has been set at V_{stress}. Therefore, the circuit only needs to put V_{stress} to the other source/drain terminal of PU2 in order to put it under stress mode.

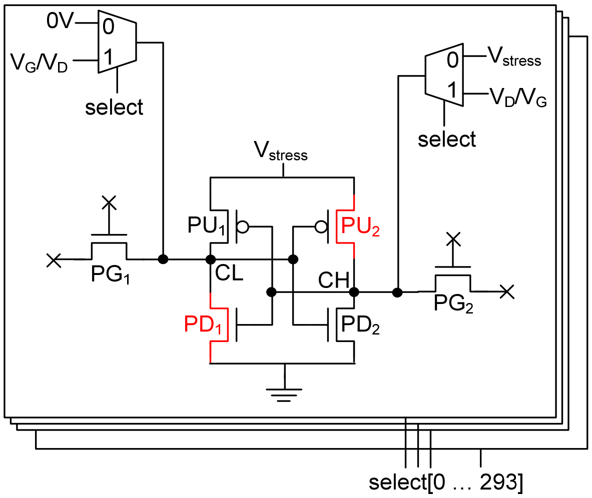


Fig. 4.21 Circuit schematic of a pair of NMOS and PMOS transistors placed under stress mode (adapted from [43]).

4.2.3 Test Setup and Measurement

The fabricated test chip was sent back in a form of a 12-inch wafer. The wafer was then diced into quarters (as in Fig. 4.22), making it manageable to be used with a Cascade probe station. A custom-made Probe card with 72 probe tips is used to probe the pads. The I/O pad configuration is shown in Fig. 4.23. Since the probe tip alignment has to be done manually, extreme care must be taken when landing the probe tips. It is easiest to first adjust the rotation by practicing landing the probe tips on nearby patterns that are just adjacent the active device area. Once the rotation is correct, the vertical and horizontal adjustments can be done on the active die without too much difficulty. Fig. 4.24 shows a die with the probe tips landed on top. The electrical signals of a probe card are brought out through a male connector pin, which can be attached into a female socket connector.

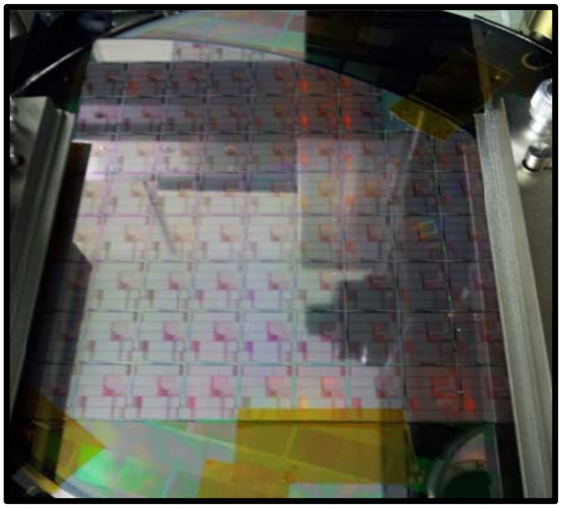


Fig. 4.22 Wafer quarter containing multiple dies.

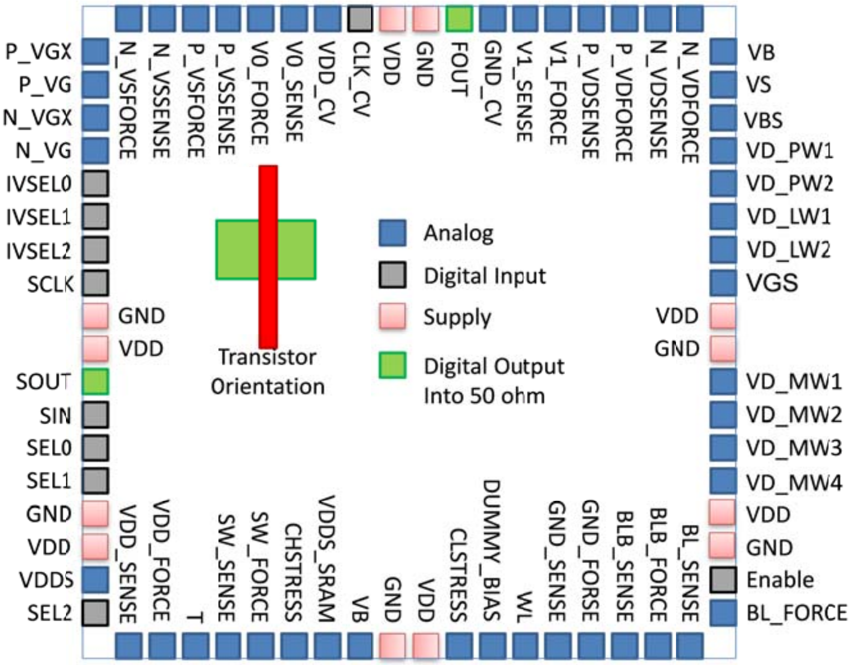


Fig. 4.23 I/O pin configuration for digital and analog signals pertaining to the SRAM array and device characterization array in the 16 nm FD-SOI test chip.

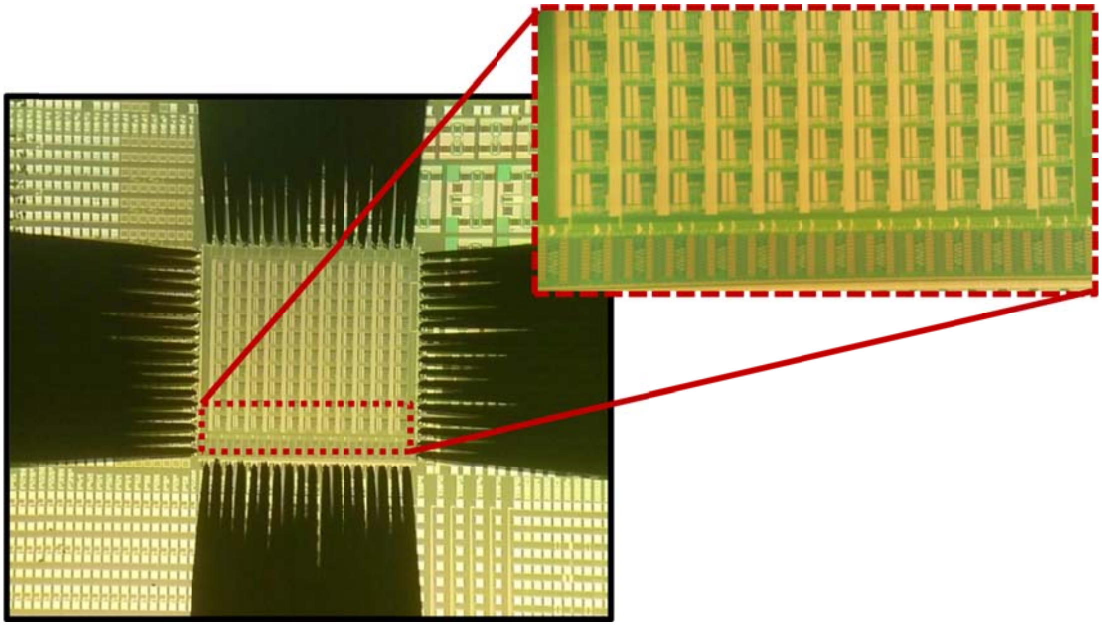


Fig. 4.24 Photo of a die under test with 72 probe tips landed on top. The inset picture shows a zoomed-in photo of the die, showing the layout of the SRAM macro located on the bottom row.

Due to the limited amount of space inside the wafer chamber of the Cascade probe station, attaching the main PCB directly to the probe card's male connector is not practical; also, its weight can result in the probe card being flexed too much. To alleviate this problem, a small breakout board is used instead to route the signals from the probe card to the main PCB board. All of the sensitive analog signals passing to the SMUs are connected through a micro-coaxial cable to fit within the small space on the breakout board. Digital signals are jumped using header pins and ribbon cables with alternating ground between the adjacent wires to help shield the signals from cross-talk and ambient electric noise. The main PCB is designed to have a DB25 connection, voltage level shifter, voltage regulator, triaxial and coaxial connectors. The test configuration including probe card, breakout board, and main PCB is shown in Fig. 4.25. A laptop computer is used to control and remotely program all characterization equipment such as the SPA or waveform generator through GPIB connections, which can be daisy-chained together. The overall test setup is depicted in Fig. 4.26.

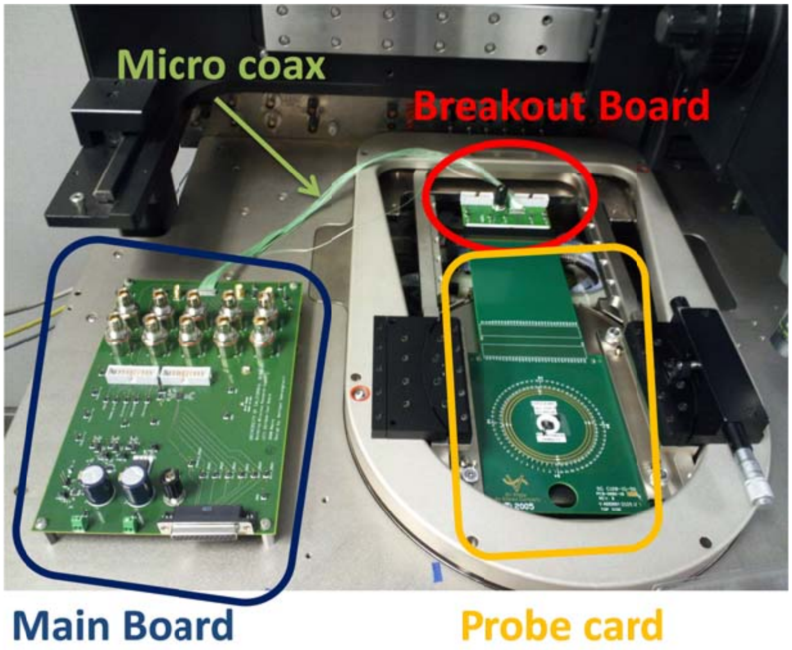


Fig. 4.25 Connections between the main PCB, breakout board, and probe card.

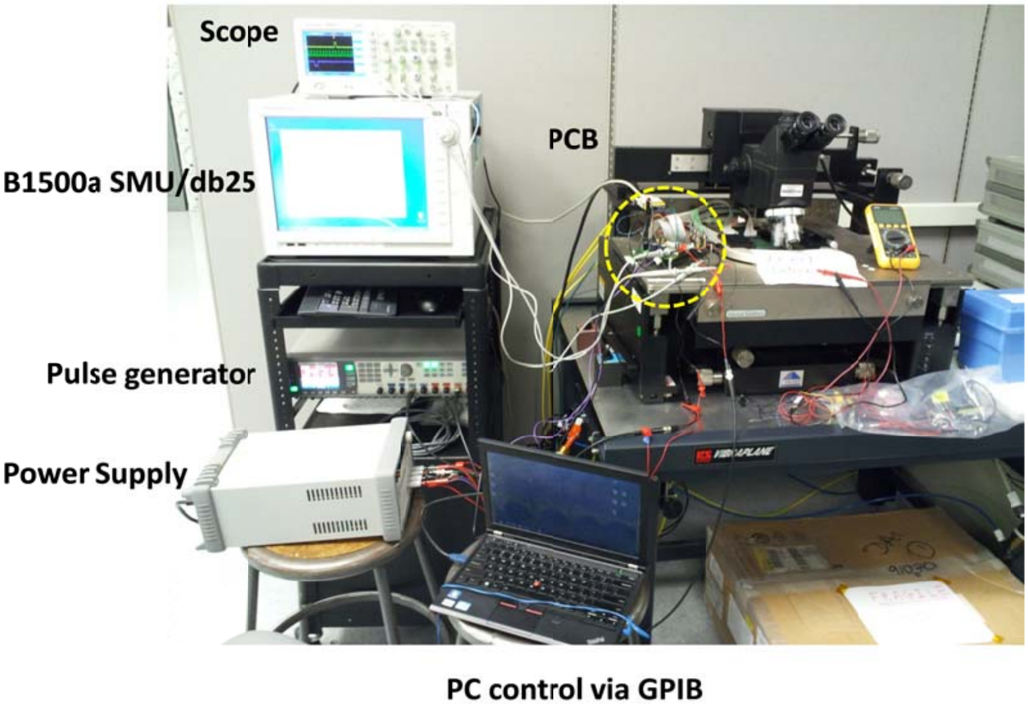


Fig. 4.26 Overall test setup consisting of semiconductor parametric analyzer (Agilent B1500A), arbitrary waveform generator (Agilent 81160), DC power supply, and a laptop computer.

The scan-chain was first tested using DB25 signals generated by the Agilent B1500. However, it was discovered that the rise time and fall time of critical signals such as S_{IN} and S_{CLK} were not short enough to give the correct operation for this test chip. As shown in Fig. 4.27, the scan-out shows an incorrect output sequence. To remedy this problem, an arbitrary waveform generator (Agilent 81160) is used instead to generate a pulse with a small rise and fall time (~50 ns). The successful scan-out operation is shown in Fig. 4.28. A robust alternative to using a pulse generator, which has only 2 output channels, is to generate all of the digital signals via a FPGA board instead.

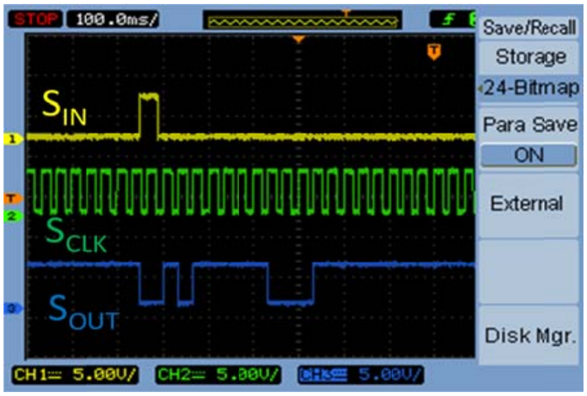


Fig. 4.27 Incorrect scan-out due to long rise and fall time of digital input signals.

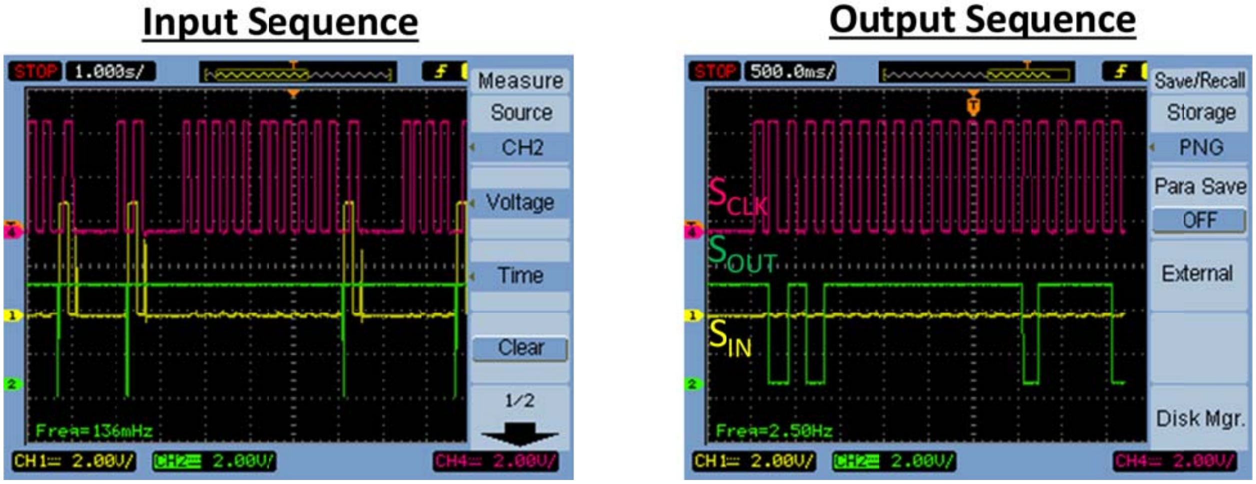


Fig. 4.28 Correct scan-out sequence when a pulse generator is used. Note that S_{OUT} is active low; the low voltage value corresponds to digital '1' logic.

4.4 Summary

The impact of device variability can be efficiently studied using a test chip vehicle. Characterization of an array of test devices including mismatch pairs, different combinations of gate length and channel width dimensions, and different layout proximity allows for the collection of data which can be used to analyze random and systematic variability. In addition to logic and analog transistors, a padded-out SRAM array is also an excellent test structure to use for characterizing the impact of variability on an actual SRAM array. Furthermore, the padded-out SRAM cell design allows one to correlate SRAM performance metrics with transistor characteristics (i.e. V_T, $I_{Effective}$, DIBL) in order to understand the root cause of the problems that affect cell yield. Since the number of I/O pads is quite limited, many signals have to be shared among test blocks either through a decoder or a multiplexer. Care must be taken when designing the selection circuitry to ensure correct operation when accessing a particular device in an array and to minimize its impact on the measured device characteristics. Leakage minimization circuits can be designed to ensure that the devices which are not under test are turned strongly off to minimize their contributions to the measured current. A switch box designed to switch between regular I-V sweep mode and voltage stress measurement mode allows for fast, built-in NBTI/PBTI and RTN characterizations - all of which are critical to study for ensuring robust SRAM operation.

4.5 References

[1] G. E. Moore, “Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff.,” *Solid-State Circuits Society Newsletter, IEEE* , vol.11, no.5, pp.33,35, Sept. 2006.

[2] S. E. Thompson, S. Parthasarathy, “Moore's law: the future of Si microelectronics,” *Materials Today*, Volume 9, Issue 6, June 2006, Pages 20-25.

[3] K. J. Kuhn, A. Murthy, R. Kotlyar, and M. Kuhn, “ Past, Present and Future: SiGe and CMOS Transistor Scaling (Invited)” *ECS Trans.* 2010 33(6): 3-17;

[4] Y. Taur; D. A. Buchanan; W. Chen; D.J. Frank; K.E. Ismail; L. Shih-Hsien; G.A, Sai-Halasz; R.G. Viswanathan ; H.-J.C. Wann; S.J. Wind; Hon-Sum Wong, “CMOS scaling into the nanometer regime,” *Proceedings of the IEEE* , vol.85, no.4, pp.486,504, Apr 1997.

[5] B Yu; C.H.J. Wann; E.D. Nowak; K. Noda; C. Hu, “Short-channel effect improved by lateral channel-engineering in deep-submicronmeter MOSFET's,” *Electron Devices, IEEE Transactions on* , vol.44, no.4, pp.627,634, Apr 1997.

[6] S.E. Thompson, P.A. Packan, M.T. Bohr, “Linear versus saturated drive current: Tradeoffs in super steep retrograde well engineering,” in *VLSI Symp. Tech. Dig.,* 1996, pp. 12 -13.

[7] K. Mistry; C. Allen; C. Auth; B. Beattie; D. Bergstrom; M. Bost; M. Brazier; M. Buehler; A. Cappellani; R. Chau; C.-H. Choi; G. Ding; K. Fischer; T. Ghani; R. Grover; W. Han; D. Hanken; M. Hattendorf; J. He; J. Hicks; R. Huessner; D. Ingerly; P. Jain; R. James; L. Jong; S. Joshi; C. Kenyon; K. Kuhn; K. Lee; H. Liu; J. Maiz; B. Mcintyre; P. Moon; J. Neirynck; S. Pae; C. Parker; D. Parsons; C. Prasad; L. Pipes; M. Prince; P. Ranade; T. Reynolds; J. Sandford; L. Shifren; J. Sebastian; J. Seiple; D. Simon; S. Sivakumar; P. Smith; C. Thomas; T. Troeger; P. Vandervoorn; S. Williams; K. Zawadzki, "A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging," *Electron Devices Meeting, 2007. IEDM 2007. IEEE International* , vol., no., pp.247,250, 10-12 Dec. 2007

[8] Y.-K. Choi; K. Asano; N. Lindert; V. Subramanian; T.-J. King; J. Bokor; C. Hu, "Ultrathin-body SOI MOSFET for deep-sub-tenth micron era," *Electron Device Letters, IEEE* , vol.21, no.5, pp.254,255, May 2000.

[9] C. Hu, "FinFET 3D Transistor & the Concept Behind it", *Solid State Techonology & Devices Seminar*, UC Berkeley, Aug 26 2011. <http://microlab.berkeley.edu/text/seminars/slides/20118_FinFET_and_the_Concept_Behind_It.pdf>

[10] X. Huang; W.-C. Lee; C. Kuo; D. Hisamoto; L. Chang; J. Kedzierski; E. Anderson; H. Takeuchi; Y.-K. Choi; K. Asano; V. Subramanian; T.-J. King; J. Bokor; C. Hu, "Sub 50-nm FinFET: PMOS," *Electron Devices Meeting, 1999. IEDM '99. Technical Digest. International*, vol., no., pp.67,70, 5-8 Dec. 1999.

[11] D. Hisamoto; W.-C. Lee; J. Kedzierski; H. Takeuchi; K. Asano; C. Kuo; E. Anderson; T.-J. King; J. Bokor; C. Hu, "FinFET-a self-aligned double-gate MOSFET scalable to 20 nm," *Electron Devices, IEEE Transactions on* , vol.47, no.12, pp.2320,2325, Dec 2000.

[12] Hisamoto, D., "FD/DG-SOI MOSFET-a viable approach to overcoming the device scaling limit," *Electron Devices Meeting, 2001. IEDM '01. Technical Digest. International* , vol., no., pp.19.3.1,19.3.4, 2-5 Dec. 2001.

[13] K. Suzuki; T. Tanaka; Y. Tosaka; H. Horie; Y. Arimoto, "Scaling theory for double-gate SOI MOSFET's," *Electron Devices, IEEE Transactions on* , vol.40, no.12, pp.2326,2329, Dec 1993.

[14] J.-W. Yang; J.G. Fossum, "On the feasibility of nanoscale triple-gate CMOS transistors," *Electron Devices, IEEE Transactions on* , vol.52, no.6, pp.1159,1164, June 2005.

[15] C. Auth; C. Allen; A. Blattner; D. Bergstrom; M. Brazier; M. Bost; M. Buehler; V. Chikarmane; T. Ghani; T. Glassman; R. Grover; W. Han; D. Hanken; M. Hattendorf; P. Hentges; R. Heussner; J. Hicks; D. Ingerly; P. Jain; S. Jaloviar; R. James; D. Jones; J. Jopling; S. Joshi; C. Kenyon; H. Liu; R. McFadden; B. Mcintyre; J. Neirynck; C. Parker; L. Pipes; I. Post; S. Pradhan; M. Prince; S. Ramey; T. Reynolds; J. Roesler; J. Sandford; J. Seiple; P. Smith; C. Thomas; D. Towner; T. Troeger; C. Weber; P. Yashar; K. Zawadzki; K. Mistry, "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors," *VLSI Technology (VLSIT), 2012 Symposium on* , vol., no., pp.131,132, 12-14 June 2012.

[16] Q. Liu; M. Vinet; J. Gimbert; N. Loubet; R. Wacquez; L. Grenouillet; Y. Le Tiec; A. Khakifirooz; T. Nagumo; K. Cheng; H. Kothari; D. Chanemougame; F. Chafik; S. Guillaumet; J. Kuss; F. Allibert; G. Tsutsui; J. Li; P. Morin; S. Mehta; R. Johnson; L.F. Edge; S. Ponoth; T. Levin; S. Kanakasabapathy; B. Haran; H. Bu; J.-L. Bataillon; O. Weber; O. Faynot; E. Josse; M. Haond; W. Kleemeier; M. Khare; T. Skotnicki; S. Luning; B. Doris; M. Celik; R. Sampson, "High performance UTBB FDSOI devices featuring 20nm gate length for 14nm node and beyond," *Electron Devices Meeting (IEDM), 2013 IEEE International* , vol., no., pp.9.2.1,9.2.4, 9-11 Dec. 2013.

[17] S.H. Rasouli; H.F. Dadgour; K. Endo; H. Koike; K. Banerjee, "Design Optimization of FinFET Domino Logic Considering the Width Quantization Property," *Electron Devices, IEEE Transactions on* , vol.57, no.11, pp.2934,2943, Nov. 2010.

[18] M. Bruel; B. Aspar; B. Charlet; C. Maleville; T. Poumeyrol; A. Soubie; A.-J. Auberton-Herve; J.M. Lamure; T. Barge; F. Metral; S. Trucchi, " 'Smart cut': a promising new SOI material technology," *SOI Conference, 1995. Proceedings., 1995 IEEE International* , vol., no., pp.178,179, 3-5 Oct 1995.

[19] A. Hokazono; S. Balasubramanian; K. Ishimaru; H. Ishiuchi; T.-J. King; C. Hu, "MOSFET design for forward body biasing scheme," *Electron Device Letters, IEEE* , vol.27, no.5, pp.387,389, May 2006.

[20] C. Fenouillet-Beranger; P. Perreau; T. Benoist; C. Richier; S. Haendler; J. Pradelle; J. Bustos; P. Brun; L. Tosti; O. Weber; F. Andrieu; B. Orlando; D. Pellissier-Tanon; F. Abbate; C. Pvichard; R. Beneyton; M. Gregoire; J. Ducote; P. Gouraud; A. Margain; C. Borowiak; R. Bianchini; N. Planes; E. Gourvest; K.K. Bourdelle; B.Y. Nguyen; T. Poiroux; T. Skotnicki; O. Faynot; F. Boeuf, "Impact of local back biasing on performance in hybrid FDSOI/bulk high-k/metal gate low power (LP) technology," *Ultimate Integration on Silicon (ULIS), 2012 13th International Conference on* , vol., no., pp.165,168, 6-7 March 2012.

[21] L.-T. Pang; B. Nikolic, “Measurement and analysis of variability in 45nm strained-Si CMOS technology,” *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE* , vol., no., pp.129,132, 21-24 Sept. 2008.

[22] J.V. Faricelli, “Layout-dependent proximity effects in deep nanoscale CMOS,” *Custom Integrated Circuits Conference (CICC), 2010 IEEE* , vol., no., pp.1,8, 19-22 Sept. 2010.

[23] M.J.M. Pelgrom, A. C. J. Duinmaijer, AP.G. Welbers, “Matching properties of MOS transistors,” *Solid-State Circuits, IEEE Journal of*, vol.24, no.5, pp.1433,1439, Oct 1989.

[24] K. J. Kuhn, M. D. Giles, D. Becher, P. Koler, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai, “Process technology variation,” IEEE Trans. Electron Devices, vol. 58, no. 8, pp. 2197–2208, Aug. 2011.

[25] A. Asenov, “Simulation of statistical variability in nano MOSFETs,” *IEEE Symp. VLSI Technol., Dig. Tech. Papers*, Jun. 2007, pp. 86–87.

[26] C.-H. Jan; U. Bhattacharya; R. Brain; S.-J. Choi; G. Curello; G. Gupta; W. Hafez; M. Jang; M. Kang; K. Komeyli; T. Leo; N. Nidhi; L. Pan; J. Park; K. Phoa; A. Rahman; C. Staus; H. Tashiro; C. Tsai; P. Vandervoorn; L. Yang; J.-Y. Yeh; P. Bai, “A 22nm SoC platform technology featuring 3-D tri-gate and high-k/metal gate, optimized for ultra low power, high performance and high density SoC applications,” *Electron Devices Meeting (IEDM), 2012 IEEE International* , vol., no., pp.3.1.1,3.1.4, 10-13 Dec. 2012.

[27] B. Pelloux-Prayer; A. Valentian; B. Giraud; Y. Thonnart; J.-P. Noel; P. Flatresse; E. Beigne, “Fine grain multi-VT co-integration methodology in UTBB FD-SOI technology,” *Very Large Scale Integration (VLSI-SoC), 2013 IFIP/IEEE 21st International Conference on* , vol., no., pp.168,173, 7-9 Oct. 2013.

[28] R-A. Bianchi; G. Bouche; O. Roux-dit-Buisson, “Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance,” *Electron Devices Meeting, 2002. IEDM '02. International* , vol., no., pp.117,120, 8-11 Dec. 2002.

[29] M.G. Bardon; V. Moroz; G. Eneman; P. Schuddinck; M. Dehan; D. Yakimets; D. Jang; G. Van der Plas; A. Mercha; A. Thean; D. Verkest; A. Steegen, “Layout-induced stress effects in 14nm & 10nm FinFETs and their impact on performance,” *VLSI Technology (VLSIT), 2013 Symposium on* , vol., no., pp.T114,T115, 11-13 June 2013.

[30] J. Davis, “Density Requirements at 28 nm”, *Design How-To*, EE Times 3/12/2012. < http://www.eetimes.com/document.asp?doc_id=1279471>

[31] M. Orshansky , S. Nassif , D. Boning, Design for Manufacturability and Statistical Design: A Constructive Approach, Springer Publishing Company, Incorporated, 2010.

[32] G. Scott, G.; J. Lutze; M. Rubin; F. Nouri; M. Manley, “NMOS drive current reduction caused by transistor layout and trench isolation induced stress,” *Electron Devices Meeting, 1999. IEDM '99. Technical Digest. International* , vol., no., pp.827,830, 5-8 Dec. 1999.

[33] P.G. Drennan; M.L. Kniffin; D.R. Locascio, “Implications of Proximity Effects for Analog Design,” *Custom Integrated Circuits Conference, 2006. CICC '06. IEEE* , vol., no., pp.169,176, 10-13 Sept. 2006.

[34] X. Sun; Q. Lu; V. Moroz; H. Takeuchi; G. Gebara; J. Wetzel; S. Ikeda; C. Shin; T.-J. King, “Tri-Gate Bulk MOSFET Design for CMOS Scaling to the End of the Roadmap,” *Electron Device Letters, IEEE* , vol.29, no.5, pp.491,493, May 2008.

[35] C. Shin; X. Sun; T.-J. King, “Study of Random-Dopant-Fluctuation (RDF) Effects for the Trigate Bulk MOSFET,” *Electron Devices, IEEE Transactions on*, vol.56, no.7, pp.1538,1542, July 2009.

[36] R.A. Vega; T.-J. King, “Comparative Study of FinFET Versus Quasi-Planar HTI MOSFET for Ultimate Scalability,” *Electron Devices, IEEE Transactions on*, vol.57, no.12, pp.3250,3256, Dec. 2010.

[37] K. Zhang, Embedded Memories for Nano-Scale VLSIs, Integrated Circuits and Systems Series, Springer, 2009.

[38] Z. Guo; A. Carlson; L.-T. Pang; K.T. Duong; T.-J. King; B. Nikolic, “Large-Scale SRAM Variability Characterization in 45 nm CMOS,” *Solid-State Circuits, IEEE Journal of* , vol.44, no.11, pp.3174,3192, Nov. 2009.

[39] A. E. Carlson, “Device and circuit techniques for reducing variation in nanoscale SRAM,” Ph.D. dissertation, UC Berkeley, May 2008.

[40] S.V. Kumar; C.H. Kim; S.S. Sapatnekar, “Impact of NBTI on SRAM read stability and design for reliability,” *Quality Electronic Design, 2006. ISQED '06. 7th International Symposium on* , vol., no., pp.6 pp.,218, 27-29 March 2006.

[41] K. Takeuchi; T. Nagumo; K. Takeda; S. Asayama; S. Yokogawa; K. Imai; Y. Hayashi, “Direct observation of RTN-induced SRAM failure by accelerated

testing and its application to product reliability assessment," *VLSI Technology (VLSIT), 2010 Symposium on* , vol., no., pp.189,190, 15-17 June 2010.

[42] S.O. Toh; T.-J. King; B. Nikolic, "Impact of random telegraph signaling noise on SRAM stability," *VLSI Technology (VLSIT), 2011 Symposium on* , vol., no., pp.204,205, 14-16 June 2011.

[43] S.O. Toh, "Nanoscale SRAM Variability and Optimization," Ph.D. dissertation talk, UC Berkeley, Mar. 2011.

Chapter 5

Variability in Germanium-Source Tunnel FETs

5.1 Introduction

A MOSFET switches on/off via the modulation of an energy barrier to thermal diffusion; therefore the steepest sub-threshold swing that it can achieve is 60 mV/dec at room temperature. This limits its on/off current ratio (I_{ON}/I_{OFF}) for low-voltage (sub-threshold) operation and hence the energy efficiency of CMOS circuitry [1]. Because of this limitation, there has been a strong push toward finding a MOSFET-replacement device, one that can achieve higher I_{ON}/I_{OFF} for a given supply voltage (V_{DD}). Of the various candidates proposed, the tunnel field-effect transistor (TFET) is emerging as one of the promising devices [2-4]. Since a TFET switches on/off via alignment/misalignment of energy bands, its minimum sub-threshold swing can be less than 60 mV/dec [5-6]. One of the challenges for the TFET to become a practical alternative to the MOSFET is its relatively low on-state drive current, which is limited by the rate of carrier tunneling. To overcome this challenge, a reduction in the effective tunneling band-gap is necessary. This can be achieved by using a smaller band-gap material in the source region. Indeed, the use of germanium (Ge) as the source material within a silicon n-channel TFET has resulted in the highest I_{ON}/I_{OFF} reported to date for a TFET operating at low voltage (0.5V) [7].

As transistor dimensions are scaled down to provide for improved performance and cost per function, random variability (*vs.* systematic variability) in transistor performance grows in significance and will present a major challenge for achieving high yield in the manufacture of integrated circuits utilizing MOSFETs with sub-30 nm gate lengths [8]. Sources of random variability include random dopant fluctuations (RDF), gate line-edge roughness (LER), and gate work function variation (WFV) [8-9]. Previous studies of variability in TFET performance have focused on systematic sources of variation [10]. In this chapter, variability in Ge-source TFET performance due to RDF is investigated via three-dimensional (3D) device simulations.

5.2 Random Dopant Fluctuation in Ge-Source TFET

5.2.1 Nominal Tunnel FET Design

Fig. 5.1a shows the 3D TFET structure, adapted from a prior design optimization study [11]. The source region is comprised of p-type germanium, while the channel region comprises p-type silicon and the drain region comprises heavily doped n-type silicon. The source doping profile is assumed to be abrupt, since it is formed by selective growth of *in-situ*-boron-doped Ge at relatively low temperature (425^{o}C) [7]. For simplicity, the drain doping profile is assumed to be abrupt and perfectly aligned to the gate edge. (For low operating voltages, gate-induced drain leakage is not significant. Also, drain-induced barrier lowering does not occur in a TFET unless the effective gate length, defined as the distance between the doping concentrations between the source and drain, is scaled aggressively [12]. Thus, the drain doping profile does not significantly impact the performance of a Ge-source TFET.) The nominal values of the geometrical device parameters defined in Fig. 1b are summarized in Table I. The nominal device width is 30 nm. A supply voltage of 0.5 V is assumed, unless otherwise stated.

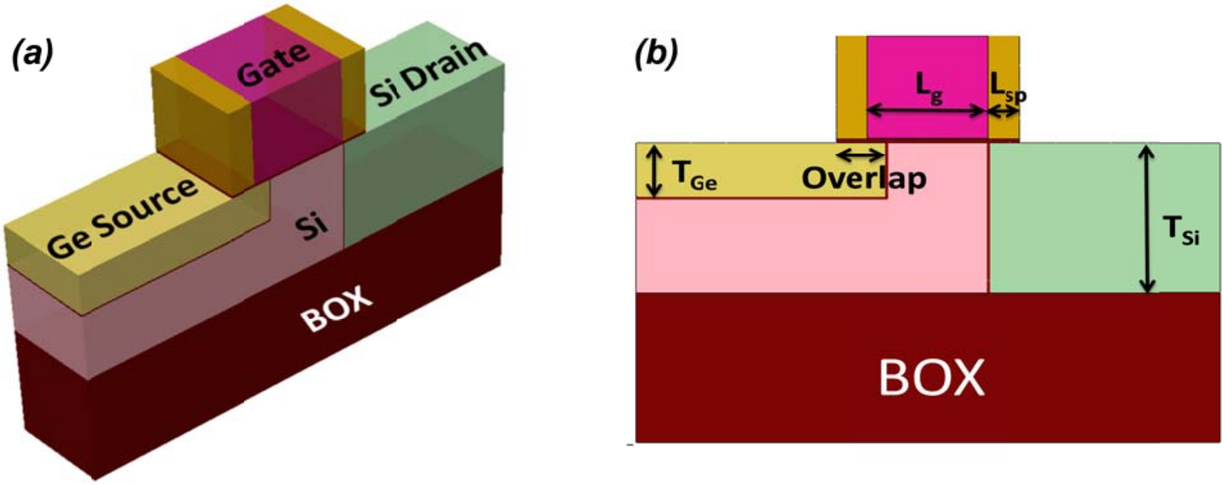


Fig. 5.1 (a) Isometric view of the Ge-source TFET structure studied in this work, (b) Cross-sectional view showing the various geometrical design parameters.

TABLE 5.1

DEVICE DIMENSIONS USED IN THIS WORK

Device Parameter	Nominal Value
Gate Length (L_g)	30 nm
Germanium Thickness (T_{Ge})	15 nm
Spacer Length (L_{sp})	8 nm
Germanium Overlap (Overlap)	13 nm
Silicon Thickness (T_{si})	40 nm
Equivalent Oxide Thickness (EOT)	1 nm
Width (W)	30 nm

The 3D device simulations were performed using Sentaurus Device [13], which uses an algorithm for dynamically determining the non-local tunneling rate, *i.e.* it calculates the band-to-band tunneling rate in multiple directions without *a priori* knowledge of the tunneling locations and vectors. The tunneling model is calibrated to experimental data for polycrystalline-Ge source TFETs (A = 1.46 x 10^{17} $cm^{-3}\cdot s^{-1}$ and B =3.59 x 10^6 $V\cdot cm^{-1}$) [14]. Polycrystalline Ge has a high density of defects with associated trap state energy level close to the valence-band edge, so that these defects effectively lower the tunnel band gap and therefore enhance I_{ON} [7] in contrast with mid-gap states which would degrade subthreshold swing and I_{OFF}.

A positive fixed charge of 8.5×10^{12} q/cm^2 at the interface between the germanium and the SiO_2 gate dielectric is assumed, as in [15]. (This was necessary to fit the device simulation to the measured I_D-V_G characteristic, and is not unreasonable considering that the SiO_2 was exposed to a dry etch process [7] prior to selective Ge growth and that the Ge-SiO_2 interface is known to be poor [16].) Carrier transport is modeled using the standard drift-diffusion models. Bandgap narrowing is modeled using Oldslotboom model. Quantum confinement effect is taken into account using Modified local-density approximation (MLDA) model.

5.2.2 Methodology for Implementing Random Dopant Fluctuations

The methodology proposed by Sano [17] is used to investigate the impact of RDF on TFET performance. Following this methodology, the randomized doping profiles are generated from a nominal structure with continuum doping profile. The dopant atom locations are randomized, and a doping function is assigned to each discrete dopant atom. (The doping function only includes the long-range portion of the Coulombic potential of the ionized dopant atom, to avoid unrealistic singularities in the potential profile [18].) The superposition of these doping functions yields the random doping concentration profile. To obtain statistically significant results, an ensemble of 200 device structures with microscopically different doping profiles were simulated for each particular TFET design. Fig. 5.2 shows one example of a TFET with randomized doping profiles in the source and channel regions.

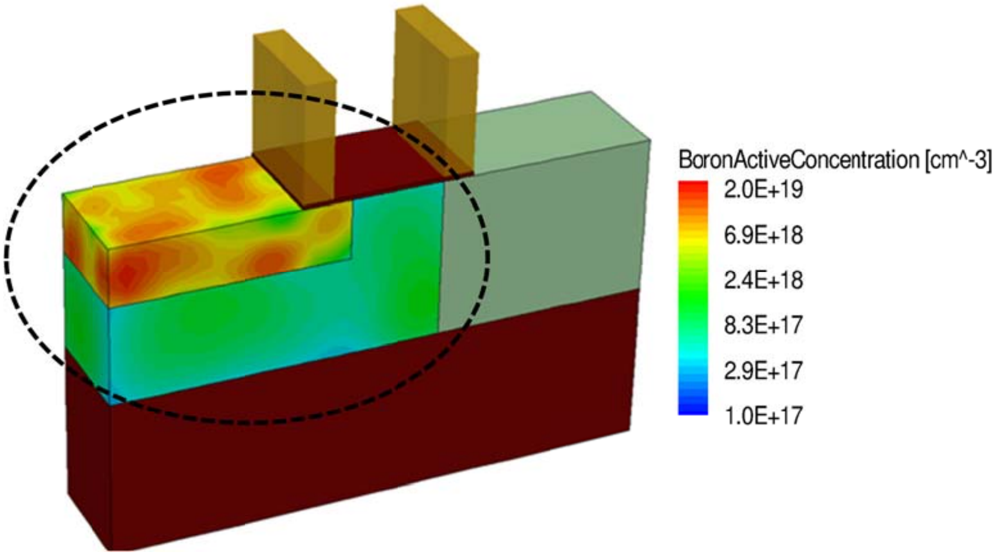


Fig. 5.2 Example of a TFET structure with randomized doping profiles in the source and channel regions. Gate electrode is omitted for clarity.

To better understand and accurately assess the impact of RDF on TFET performance, the randomization algorithm is selectively applied to the different regions of the device. In addition, RDF-induced variations are studied for different combinations of nominal doping concentrations within the source and channel regions, to assess the tradeoff between nominal device performance and variability.

5.2.3 Modeling Limitations

The commercial TCAD package used in this work offers a robust and efficient way to study atomistic effects in semiconductor device structures, so that statistical results for TFETs of different designs can be generated without computationally expensive quantum mechanical simulations. The approximations used in Sano's algorithm and the band-to-band tunneling model each may give rise to significant errors. As mentioned above, Sano's algorithm considers only the long-range Coulombic potential profile of each discrete dopant atom; it does not include the short-range Coulombic potential profile, to avoid singularities which can introduce artifacts in the conventional drift-diffusion simulator. The band-to-band tunneling model used in this work is an extension of the Kane and Keldysh model to arbitrary energy-band profiles; it searches for the most probable straight-line tunneling path and calculates the corresponding tunneling barrier. The tunneling energy is equal to the valence band energy at the starting position and it is also equal to the conduction band energy (plus any band offset) at the ending position. A tunneling path undergoes specular reflection when it encounters Neumann boundaries. [13]

It is not clear how the modeled tunneling current would be affected by the short-range Coulombic potentials of discrete dopant atoms. Sano's approach uses a screening parameter that is doping-dependent to distinguish the short- and long-range portions of the Coulombic potential profile. It is possible that additional screening by carriers induced by the gate in the transistor on state could affect the tunneling current. An *ab initio* study is necessary to elucidate the extent of such quantum mechanical effects.

The simulation approach taken in this study is not intended to precisely account for every subtle physical phenomenon. Rather, it uses established analytical models carefully calibrated to experimental data to reasonably approximate physical processes. (Note: the A and B coefficient values may be different for different doping configurations and under the influence of the short-range Coulombic potential of the discrete dopant atoms. Because of limited reliable experimental data for TFETs reported in the literature, a comprehensive calibration for all possible doping configurations is not possible at the present time.) The limitations of this approach are a subject of ongoing research from both theoretical and experimental perspectives. Thus, the reader should keep in mind that the findings reported herein should be viewed qualitatively rather than quantitatively.

5.2.4 Impact of RDF on V_{TH} Variation for Optimized Nominal Design

A previous design optimization study [11] found that a source doping concentration of $N_S = 10^{19}$ cm^{-3}, channel doping concentration of $N_{CH} = 10^{18}$ cm^{-3}, and drain doping concentration of $N_D = 10^{19}$ cm^{-3} provides for maximum I_{ON}/I_{OFF} for a vertical tunneling TFET design, wherein tunneling occurs primarily within the Ge source. This particular design serves as the starting point for the current study. To distinguish the variability contribution from each region of the device, the doping profiles within the source, channel, and drain regions are randomized separately, as well as together. Fig. 5.3a shows the simulated I_D-V_G curves for different randomized doping profiles in each of the source, channel, and drain regions. Similarly as for the MOSFET [8], the average threshold voltage of a TFET is reduced with randomized doping profiles. The transistor can be considered as many narrow transistors connected in parallel, each narrow transistor comprising one slice of the transistor. In a TFET, V_{TH} is largely set by the slice in which band-to-band tunneling occurs first (*i.e.* at the lowest gate voltage). Thus, all it takes for V_{TH} lowering to occur is the presence of a few such slices. The probability of finding those few slices with lower V_{TH} compared to the nominal V_{TH} is greater than finding all the slices with larger V_{TH}; therefore RDF is more likely to lower V_{TH} than to raise V_{TH}. To elucidate how RDF affects V_{TH}, the randomized doping profiles for the devices with the highest and lowest V_{TH} values in Fig. 5.3a are shown in Fig. 5.3b. It can be seen that, for this particular TFET design, high V_{TH} corresponds to a device with higher dopant concentration in the gate-to-source overlap region while low V_{TH} corresponds to a device with lighter dopant concentration in this region. This is reasonable since a larger voltage drop is required to invert the surface of a more heavily doped Ge source.

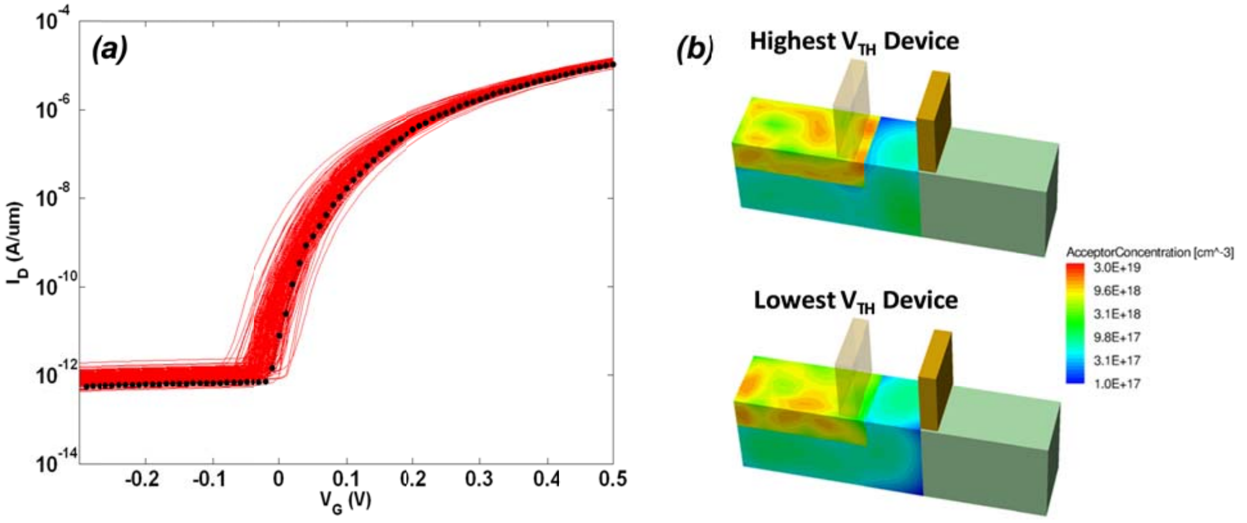


Fig. 5.3 (a) Simulated TFET I_D – V_G curves for 200 microscopically different (randomized) doping profiles, for nominal doping concentrations $N_S = 10^{19}$ cm^{-3}, $N_{CH} = 10^{18}$ cm^{-3}, and $N_D = 10^{19}$ cm^{-3}. The black symbols show the simulated I_D-V_G curve for continuum doping profiles, as reference. V_{DS} = 0.5 V. Gate work function WF = 4.0 eV, (b) Randomized doping profiles for the devices with the highest and lowest V_{TH} values in Fig. 5.3a. Gate electrode is omitted for clarity.

Note that I_{OFF} is relatively insensitive to V_G below a certain point. In this range of gate voltages, the dominant component of current is due to p-n junction leakage. To maximize the average I_{ON}/I_{OFF}, the gate work function should be tuned so that the onset of band-to-band tunneling occurs at 0 V for the device with the lowest V_{TH}. In this manner, variation in I_{OFF} is minimized.

The definition of the threshold voltage (V_{TH}) for a tunnel FET is still in debate [19]. In this study, V_{TH} is defined to be the gate voltage corresponding to a drain current of 1 nA/um, for $V_{DS} = V_{DD} = 0.5$V. The V_{TH} variations resulting from RDF in different regions of the device are summarized in Fig. 5.4. σV_{TH} due to RDF in the source region (13.25 mV) accounts for ~95% of the σV_{TH} due to RDF in all regions (14 mV). This indicates that the source's contribution to random V_{TH} variation is the largest, followed by the channel's contribution and the drain's contribution. If the contributions from the different regions are independent of one another, the overall σV_{TH} can be calculated from individual contributions according to the equation:

$$\sigma V_{TH}|_{Overall} \approx \sqrt{(\sigma V_{TH})^2|_{Source} + (\sigma V_{TH})^2|_{Channel} + (\sigma V_{TH})^2|_{Drain}} \qquad (5.1)$$

$\sigma V_{TH}|_{Overall}$ is calculated to be 13.58 mV, which is in close agreement with the value obtained through device simulation.

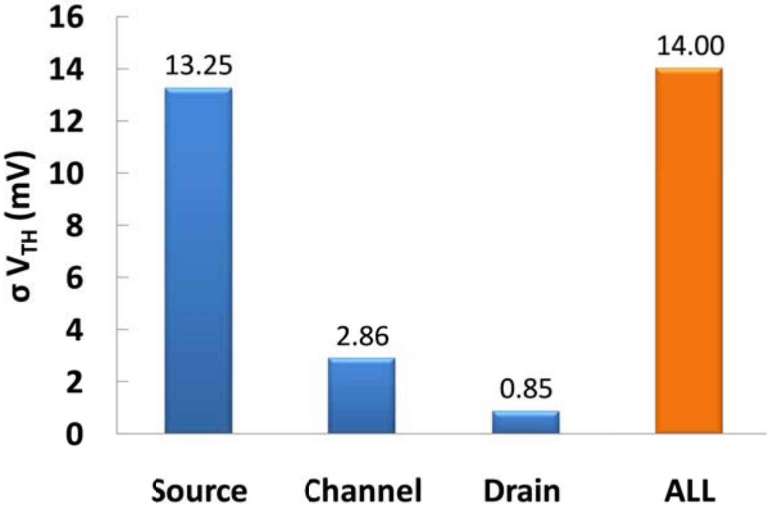


Fig. 5.4 RDF-induced contributions to σV_{TH} for each region of the TFET. σV_{TH} due to RDF in all of the regions is also shown.

5.2.5 Impact of RDF for Various Nominal Source Doping Concentrations

Since RDF in the source region accounts for most of the variability in V_{TH} of an optimally designed TFET, it is worthwhile to investigate the impact of the nominal source doping concentration. For this investigation, the channel and drain doping concentrations are fixed at $N_{CH} = 10^{18}$ cm^{-3} and $N_D = 10^{19}$ cm^{-3}, respectively, as before. The doping randomization algorithm is applied to all the regions of the device. Fig. 5.5a shows the simulated I_D-V_G plots and Fig. 5.5b compares the σV_{TH} values, for each nominal source doping concentration N_S. σV_{TH} is minimized for $N_S = 5\times10^{19}$ cm^{-3}. The non-monotonic dependence of σV_{TH} on N_S is likely due to the change in the dominant tunneling pathway with changing N_S: as N_S increases, the dominant tunneling changes from occurring "vertically" within the Ge source to occurring "laterally" from the p-type Ge source to the n-type silicon inversion layer in the channel region. (Note that the vertical tunneling TFET design exhibits steeper local subthreshold swing compared to the lateral tunneling TFET design.) For lateral tunneling, V_{TH} is affected more by the channel doping concentration; therefore, since N_{CH} is much lower than N_S, RDF-induced V_{TH} variation is lower. For very high N_S, the area of the tunneling region becomes even smaller [20] and hence it is likely to be more sensitive to RDF, which could explain the slight increase in σV_{TH} for $N_S = 10^{20}$ cm^{-3}.

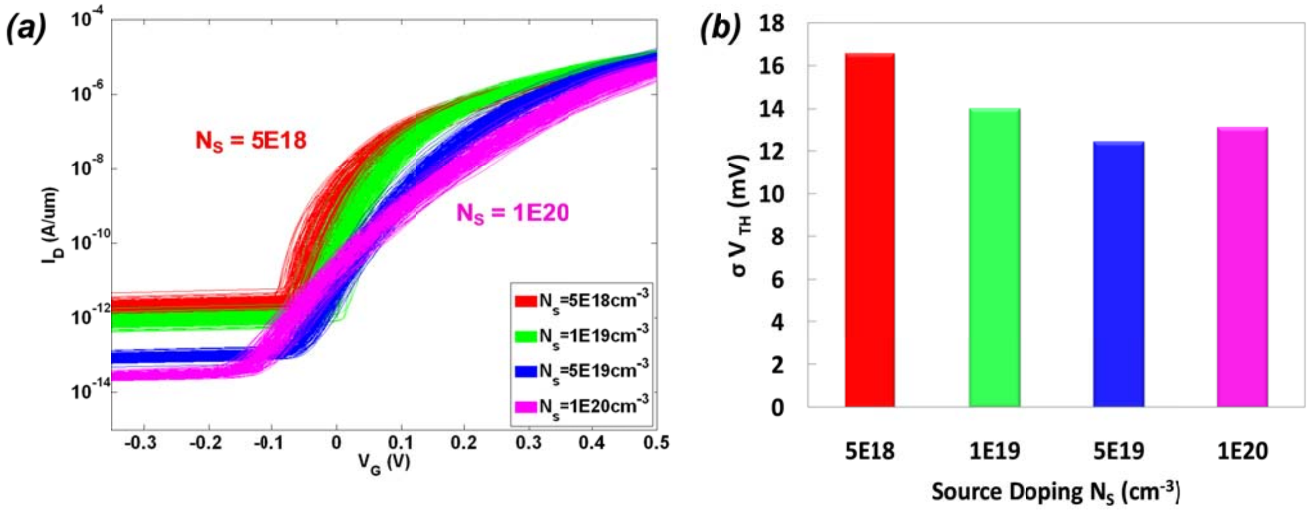


Fig. 5.5 (a) Simulated I_D-V_G curves for each nominal source doping concentration. V_{DS} = 0.5 V. Gate work function WF = 4.0 eV, (b) Comparison of σV_{TH} for different nominal source doping concentrations.

5.2.6 Impact of RDF for Various Nominal Channel Doping Concentrations

Since RDF in the channel region can significantly affect V_{TH} variation in the case of a heavily doped source, the impact of the nominal channel doping concentration is examined in this section. Specifically, two levels of nominal channel doping (light: $N_{CH} = 10^{15}$ cm^{-3}; heavy: $N_{CH} = 10^{18}$ cm^{-3}) are investigated, for the predominantly vertical tunneling design ($N_S = 10^{19}$ cm^{-3}) and the predominantly lateral tunneling design ($N_S = 10^{20}$ cm^{-3}). The doping randomization algorithm is either applied to the source region only, the channel region only, or to all of the regions. (The case where dopants are randomized only within the drain region is not presented herein, because its σV_{TH} contribution was found to be relatively small.) The RDF-induced σV_{TH} values are compared in Fig. 5.6a for the various TFET designs (different combinations of N_S and N_{CH}). RDF in the source region is the dominant source of V_{TH} variation except when the channel is heavily doped for the lateral tunneling design ($N_S = 10^{20}$ cm^{-3} and $N_{CH} = 10^{18}$ cm^{-3}), in which case RDF in the channel region becomes equally significant.

It should be noted that although the individual contributions of RDF-induced variation from different regions are largely independent of one another for the optimized TFET design (in which tunneling occurs predominantly within the Ge source), this is not the case if tunneling occurs predominantly from the source to the channel. This is evident in Fig. 5.6a, especially for the case of $N_S = 10^{20}$ cm^{-3} and $N_{CH} = 10^{18}$ cm^{-3}, and is due to interactive effects of the local source doping concentration and the local channel doping concentration for lateral tunneling.

Fig. 5.6b compares the results for the vertical tunneling design ($N_S = 10^{19}$ cm^{-3}) and the lateral tunneling design ($N_S = 10^{20}$ cm^{-3}), for randomized doping profiles in all of the TFET regions. It can be seen that the impact of increasing N_{CH} is opposite for these two designs. For the vertical tunneling design which turns on when the surface of the Ge source region becomes inverted, higher N_{CH} is desirable because it results in less depletion of the p-type Ge source by the p-type Si channel/body [11]. As a result, tunneling occurs in a direction that is more vertical, across a shorter depletion distance. The lower amount of associated depletion charge within the source region results in smaller σV_{TH}, similarly as for a MOSFET with lower channel/body doping concentration. For the lateral tunneling design which turns on when the surface of the Si channel region becomes inverted (as in a MOSFET), lower N_{CH} results in smaller depletion charge in the channel region and hence smaller σV_{TH}. Fig. 5.6b also compares the values of local subthreshold swing (SS) at $V_{GS} = V_{TH}$, extracted from simulated I_D-V_G curves for TFETs with continuum doping profiles. It is interesting to note that steeper local SS does not necessarily provide for smaller σV_{TH}. A more meaningful metric is the 'effective' subthreshold swing, which is defined as the inverse slope of the line connecting the operating I_{ON} and I_{OFF} on a $\log(I_D)$-V_G plot [21]. Therefore, for TFET design optimization, it is imperative to examine I_{ON} and I_{OFF} more closely.

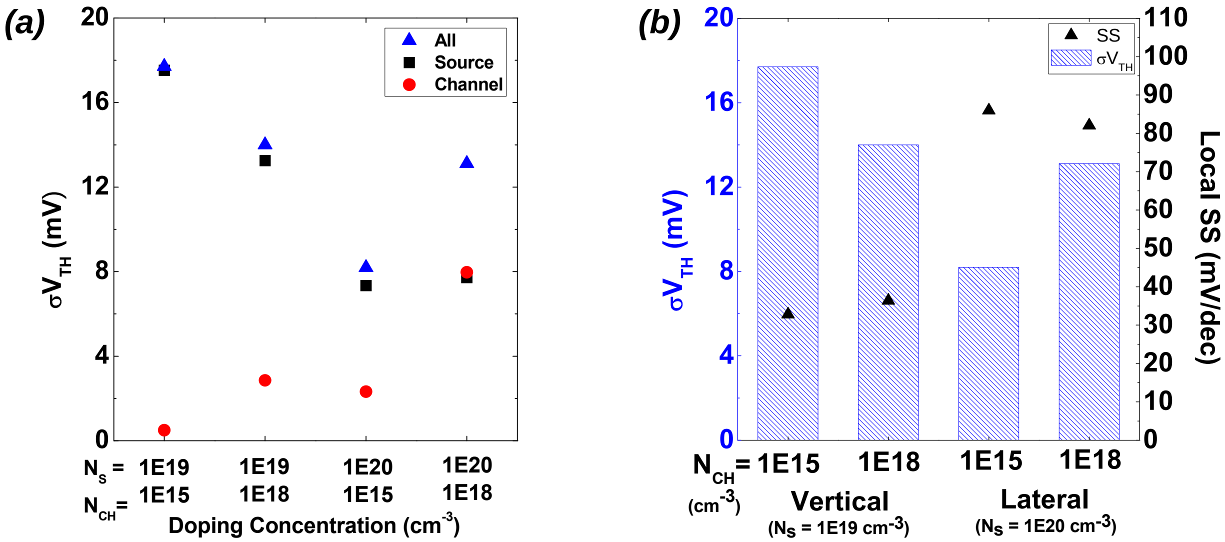


Fig. 5.6 (a) σV_{TH} contributions due to RDF in different regions of the TFET, for different combinations of source and channel nominal doping concentrations, (b) σV_{TH} resulting from RDF in all regions of the TFET, for the vertical tunneling design ($N_S = 10^{19}$ cm^{-3}) and the lateral tunneling design ($N_S = 10^{20}$ cm^{-3}). Local subthreshold swing, extracted from simulated I_D-V_G curves for TFETs with continuum doping profiles at $V_{GS} = V_{TH,}$ is also shown for reference.

5.2.7 Impact of Variability on $I_{ON} - I_{OFF}$

From the I_D-V_G curves in Fig. 5.3a and Fig. 5.5a, it can be seen that the effect of RDF is not always a simple V_{TH} shift, *i.e.* the turn-on voltage (corresponding to the onset of band-to-band tunneling) and the switching steepness each can be affected as well. Thus, from a circuit design perspective, it is also important to examine the variations in I_{ON} and I_{OFF}. Here I_{OFF} is defined as the drain current at V_{GS} = 0V, $V_{DS} = V_{DD}$ = 0.5 V, and I_{ON} is defined as the drain current at $V_{GS} = V_{DS} = V_{DD} = 0.5$ V. The gate work function is adjusted in the range from 4.0 eV to 4.35 eV to adjust I_{OFF}.

Fig. 5.7a compares the tradeoff between average I_{ON} and average $\log(I_{OFF})$ for various lateral tunneling TFET designs. $\text{Log}(I_{OFF})$ is chosen over I_{OFF} because the logarithmic of I_{OFF} has a more Gaussian-like distribution since I_{OFF} is proportional to the exponential of $V_{TH,}$ allowing for a more meaningful assessment when standard variation is taken into account. The curves for the average values are plotted with open symbols, whereas the curves with variation (3σ) taken into account are plotted with filled symbols. Without accounting for the impact of RDF-induced variations, one would conclude that for I_{ON} in the range $<10^{-6}$ A/um, $N_S = 10^{20}$ cm^{-3} and $N_{CH} = 10^{18}$ cm^{-3} is the superior design because it achieves the lowest I_{OFF} for a given I_{ON}. Due to larger variation in I_{ON} for this design, however, this turns out to be the worst design. Accounting for the impact of RDF-induced variations, one can see that $N_S = 5\times10^{19}$ cm^{-3} and $N_{CH} = 10^{18}$ cm^{-3} is the superior design. Only for applications that require very low I_{ON} ($< 10^{-8}$ A/um) would the design with $N_S = 10^{20}$ cm^{-3} and $N_{CH} = 10^{15}$ cm^{-3} be preferred.

Fig. 5.7b compares the tradeoff between average I_{ON} and average $\log(I_{OFF})$ for the best lateral tunneling TFET design ($N_S = 5\times10^{19}$ cm^{-3}, $N_{CH} = 10^{18}$ cm^{-3}) against that for the best vertical tunneling TFET design ($N_S = 10^{19}$ cm^{-3}, $N_{CH} = 10^{18}$ cm^{-3}). Accounting for the impact of RDF-induced variations, one can see that the vertical tunneling TFET design is best for I_{ON} in the range >3uA/um. This is primarily due to the smaller I_{ON} variation of the vertical tunneling design compared to the lateral tunneling design. (From Fig. 5.3a and Fig. 5.5a, one can see that the on-state current is much less sensitive to gate voltage for the vertical tunneling TFET design in comparison to the lateral tunneling TFET design.)

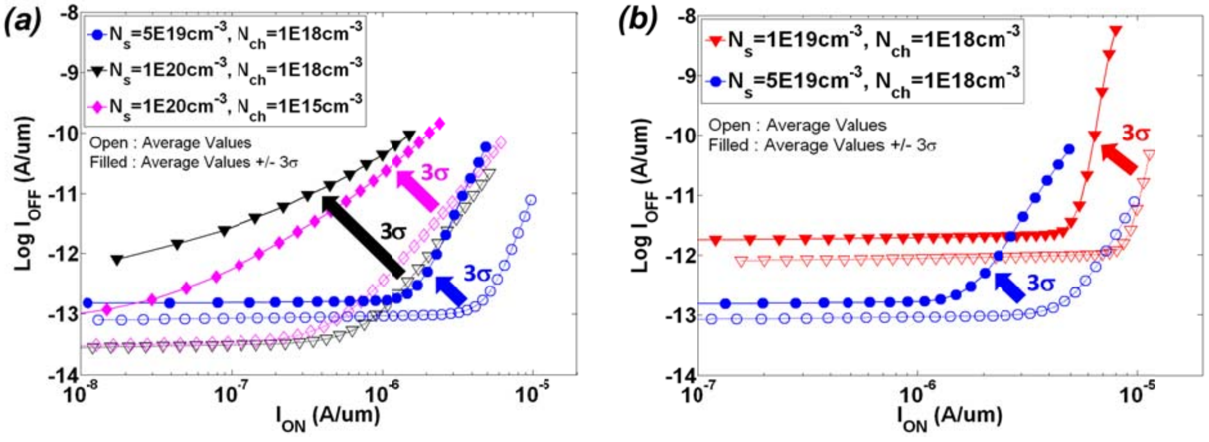


Fig. 5.7 Comparison of I_{ON}-I_{OFF} tradeoff (a) for lateral tunneling TFET designs, (b) for optimal lateral tunneling TFET design vs. optimal vertical tunneling TFET design.

5.2.8 Width Dependence of σV_{TH}

RDF-induced variation in V_{TH} mismatch (ΔV_{TH}) scales with the inverse square root of the channel width (W) for a MOSFET [22]. Since the lateral tunneling TFET turns on when the surface of the Si channel region becomes inverted, similarly to a MOSFET, it should exhibit a similar dependence of $\sigma\Delta V_{TH}$ on W. The results shown in Fig. 5.8 confirm this to be the case. The A_{Vt} values extracted from the best (least-squares) linear fit are lower than reported for advanced MOSFET structures, ~1 mV·um [23]. This may be due to the fact that tunneling depends both on the local electric potential at the point of hole generation and the local electric potential at the point of electron generation; since these two points are spatially separated (Fig. 9), there is an averaging effect which results in reduced variation (and hence mismatch) in V_{TH}. (In contrast, diffusion depends only on the local electric potential at the point of thermionic emission in a MOSFET.)

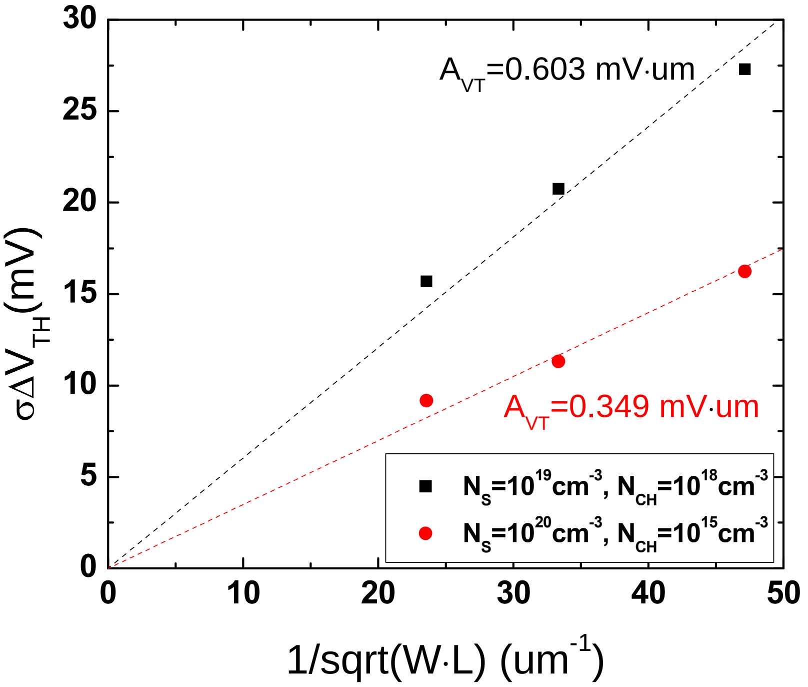


Fig. 5.8 Width dependence of σV_{TH}. The dashed lines indicate the best linear fit through the origin. V_{TH} is measured for V_{DS} = 0.5 V.

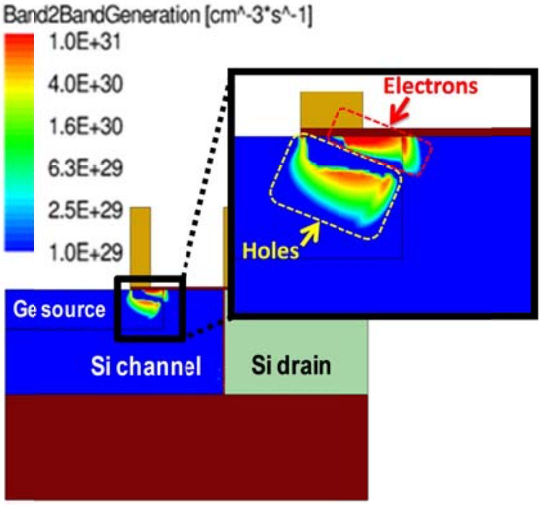


Fig. 5.9 Cross-sectional view showing contour plots for hole generation and electron generation.

5.2.9 Impact of V_{DD} Scaling

MOSFET operation at lower V_{DD} is beneficial for reducing the effect of drain-induced barrier lowering (DIBL) and hence lower σV_{TH}. That is one of the reasons why for a conventional MOSFET the $\sigma V_{TH\ Lin}$ in linear regime (V_{DS} = 50mV) is considerably smaller as compared to $\sigma V_{TH\ Sat}$ in the saturation regime (V_{DS} = 1 V). Since DIBL is not a serious issue for TFETs of this size (30 nm L_g), the benefit of V_{DD} scaling for reducing σV_{TH} should be relatively small or negligible. The results in Fig. 5.10 confirm this to be the case. The small difference in σV_{TH} values for different V_{DD} operation is most likely due to the fact that the number of instances simulated for each case is relatively small number (200).

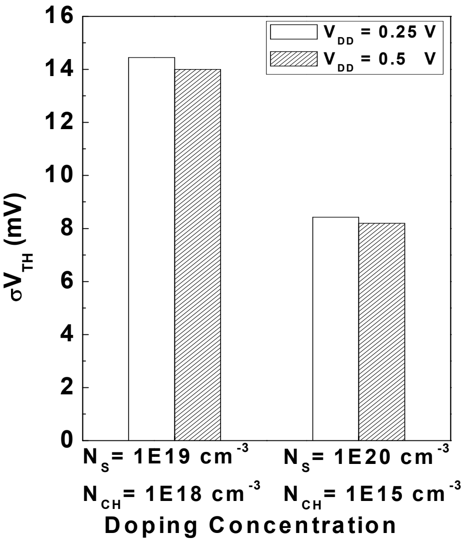


Fig. 5.10 Impact of V_{DD} scaling on σV_{TH}. ($V_{DS} = V_{DD}$.)

5.3 Line-Edge Roughness (LER) in Ge-Source TFET

In addition to random dopant fluctuation effects, as transistor gate lengths are scaled down to 30 nm and below, gate line edge roughness (LER) due to lithography and etching steps can become a significant fraction of the nominal gate length, making it a prominent source of variation in transistor performance [24],[25]. For the conventional planar bulk MOSFET, threshold voltage (V_T) variation due to gate LER may be significant compared to RDF [8].

The planar Ge-source TFET is designed for vertical (perpendicular to the semiconductor/gate-dielectric interface) tunneling in which BTBT occurs primarily within the region of the source overlapped by the gate electrode. Therefore, the electrical behavior is strongly dependent on the geometry as well as the doping of this region. The impact of RDF was previously examined in [26]. Since gate LER can affect the length of the gate-to-source overlap region, it is important to also assess the impact of gate LER. The contribution of gate LER to variability in planar Ge-source TFET performance is assessed and compared to that of RDF, using technology computer aided design (TCAD) tools to model three-dimensional (3D) device performance.

5.3.1 LER Implementation

Scanning Electron Micrograph (SEM) measurements of extreme-ultraviolet (EUV) resist lines with root mean square roughness of 3.96 nm and correlation length of 21.6 nm are used to define the gate line edge profiles. The gate-sidewall spacers are assumed to be perfectly conformal, so that their outer edges have roughness that is perfectly correlated to the gate LER. Since the Ge source region is formed by first recessing the silicon on the source side and then selectively depositing Ge to refill the etched region [7], the source (channel) edge profile can be rough. Two extreme cases of the source edge profile are considered herein: 1) “Smooth edge” case, wherein the interface between the Ge source region and the Si channel region is smooth, and 2) “Rough edge” case, wherein the Ge-source/Si-channel interface has roughness that is perfectly correlated with that of the outer edges of the gate-sidewall spacers. The nominal simulated device structure is shown in Fig. 5.11 (a), and the structures with two cases of source edge profile are shown in Fig. 5.11 (b),(c).

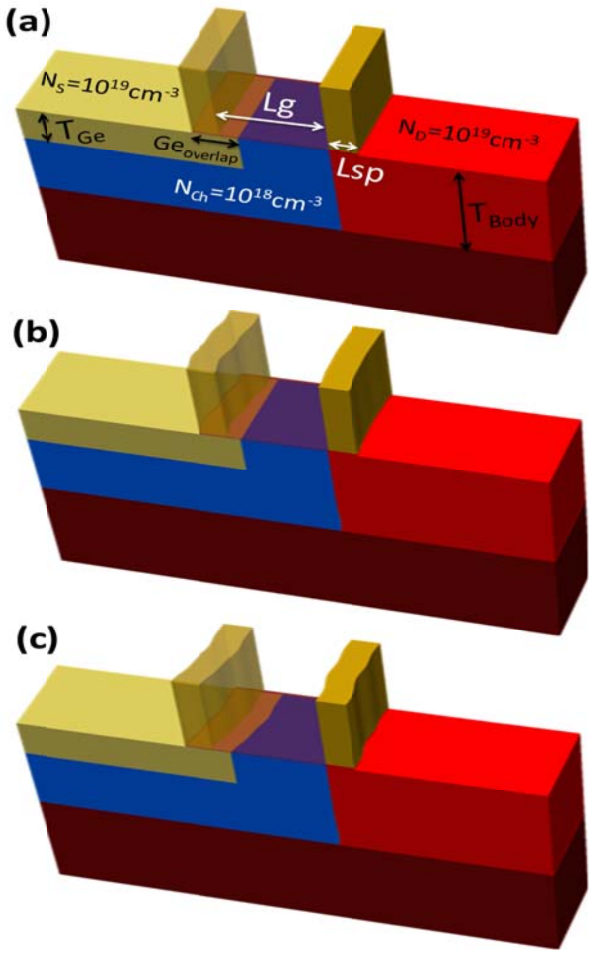


Fig. 5.11 a) Nominal 3D TFET design. Examples of TFET structures with gate LER and a Ge source region with b) smooth and c) rough interface with the Si channel region. The gate electrode is not shown for clarity.

5.3.2 Impact of LER on TFET Performance

To examine the effect of LER on TFET performance, 200 3D TFET structures with unique gate-LER profiles were generated for the Smooth edge case and the Rough edge case. The ensembles of simulated transfer TFET characteristics for the two cases are shown in Fig. 5.12.

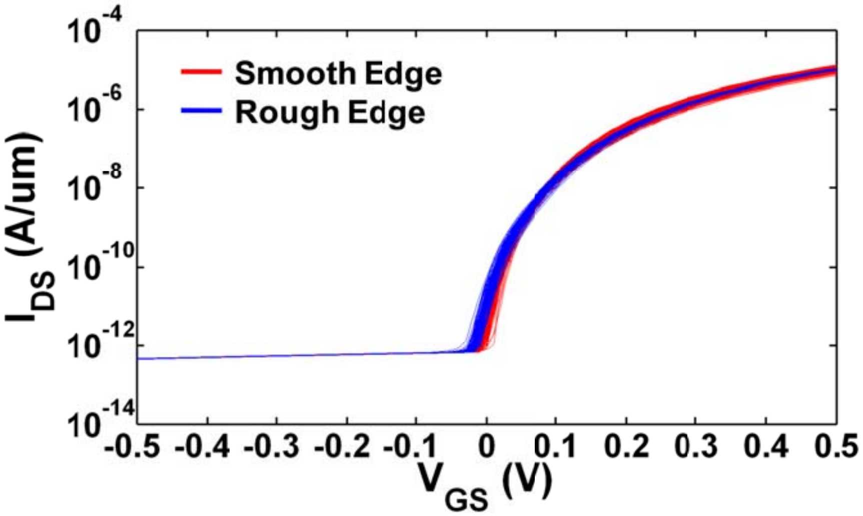


Fig. 5.12 Ensembles of simulated I_{DS}-V_{GS} characteristics for planar Ge-source TFETs with gate LER: (red) Smooth edge case and (blue) Rough edge case. 200 individual devices were simulated for each case. V_{DS} = 0.5 V. The gate work function is set to 4.0 eV.

From Fig. 5.12, one can observe that there is a gate-voltage range (V_{GS} < -0.1 V) in which the drain current is low and approximately constant, *i.e.* it reaches a "floor." In this voltage range the current is simply the leakage current of the reverse-biased p-n drain-source junction. Variation in the leakage floor current ($I_{leakage}$) induced by LER is small for both the Smooth edge case ($\sigma I_{leakage\ floor,Smooth} = 6.24\times10^{-15}$ A/um) and the Rough edge case ($\sigma I_{leakage\ floor,Rough} = 6.09\times10^{-15}$ A/um), compared to that induced by RDF ($\sigma I_{leakagefloor,RDF} = 2.68\times10^{-13}$ A/um). Additionally, the average value of $I_{leakage}$ is 1.5× higher with RDF *vs.* LER. To explain this, the energy band diagrams along the source, channel and drain regions at a distance 2 nm below the gate oxide interface are compared in Fig. 5.13 for the nominal device, a device with gate LER, and a device with RDF. It can be seen that the effect of LER is to vary the length of the channel region. Since this variation is small compared to the minority carrier diffusion lengths (several microns) within the p and n regions, variation in $I_{leakage}$ due to LER is small. In contrast, the effect of RDF is to modulate the local dopant concentration within the p and n regions – reflected by the change in potential barrier height – hence the minority carrier diffusion lengths, resulting in larger variation in $I_{leakage}$.

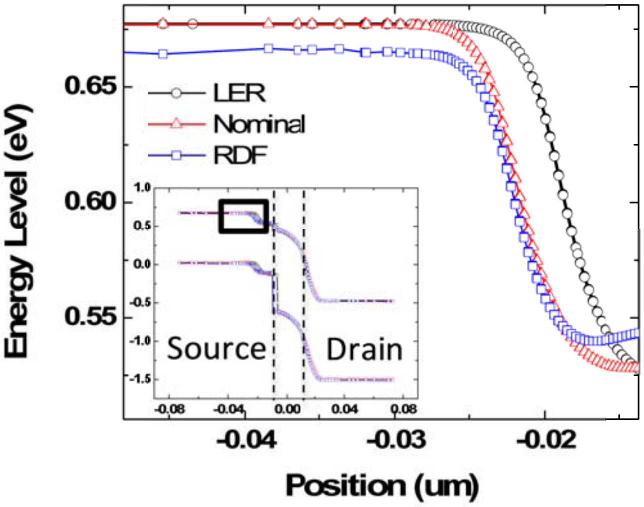


Fig. 5.13 Comparison of electron potential profiles within the source region, at a distance 2 nm below the gate oxide interface, for the nominal TFET design compared to devices with LER or RDF. Inset shows the full energy band diagrams from the source region to the drain region. Positions of the source and drain junctions for the nominal TFET are indicated.

The threshold voltage (V_T) is taken to be the gate voltage corresponding to a drain current of 10^{-9} A/um with V_{DS} = 0.5 V. The value of the standard deviation in V_T for each of the two LER cases are summarized in Table 5.2, along with that for the case of RDF only. It can be seen that V_T variation induced by LER is similar for the two source edge cases – the small difference (<1 mV) is likely due to the limited sample size (200 devices) – and is notably smaller than that due to RDF.

TABLE 5.2
VALUES OF STANDARD DEVIATION IN V_T RESULTING FROM GATE LER.

	LER Smooth Edge	LER Rough Edge	RDF
σV_T (mV)	**2.82**	**3.57**	**13.99**

To elucidate the mechanism of V_T variation induced by LER, the structures having the lowest and highest V_T values are examined in Fig. 5.14. For the Rough edge case, variation in V_T can be seen to be due to variation in the effective channel length (L_{eff}), which is defined as the shortest distance between the p+ doping in the source and the n+ doping in the drain. The low-V_T device has smaller L_{eff} and thus more influence of the drain voltage on the electric field at the source tunneling barrier (hence lower V_T), as compared to the high-V_T device.

For the Smooth edge case, the difference in V_T is primarily caused by a difference in the effective tunneling area, which corresponds to the area of overlap between the gate and the Ge source for this Ge-source TFET design [21],[22],[29]. The effective tunneling area and hence the tunneling current of the low-V_T device is larger than that of the high-V_T device, so that the threshold current level is reached at a smaller applied gate voltage. In addition to the effective tunneling area, L_{eff} also affects V_T for the Smooth edge case.

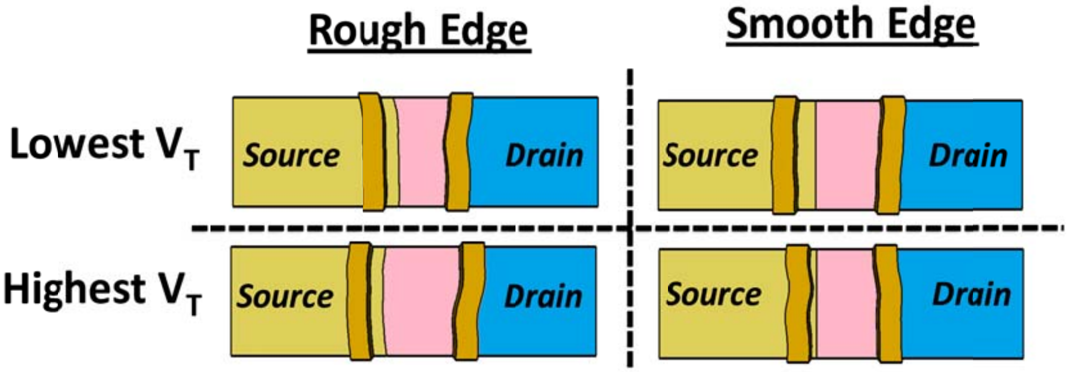


Fig. 5.14 Plan views of the TFET structures with the lowest and highest values of V_T, for each of the Rough edge and Smooth edge LER cases. The gate electrode is not shown to allow the source junction to be seen.

From Fig. 5.12 it can be seen that a rough source edge generally results in the onset of tunneling at a lower gate voltage. This can be attributed to an enhancement in the peak local electric field along the edge of a rough source, as shown in Fig. 5.15.

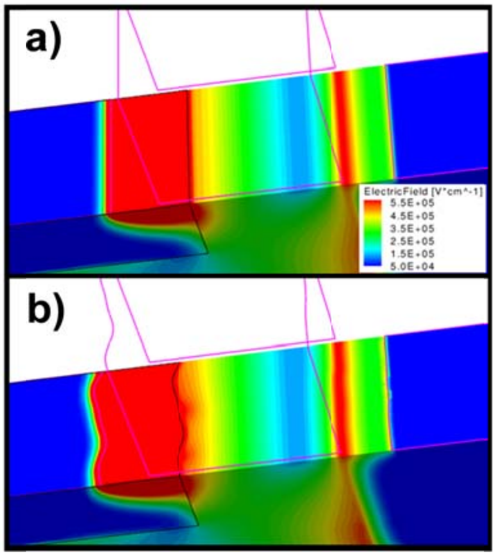


Fig. 5.15 Electric field contour plots for a) the nominal TFET structure and b) a TFET structure with gate LER and Rough source edge. V_{GS} = 0V, V_{DS} = 0.5 V. A larger peak magnitude is seen along the rough source interface as compared to the smooth source interface. Both figures are plotted using the same scale.

From Fig. 5.12 it also can be seen that a smooth source edge results in significantly greater LER-induced variation in TFET I_{ON}: σI_{ON} = 1.25x10^{-6} A/um for the Smooth edge case, while σI_{ON} = 1.86x10^{-7} A/um for the Rough edge case. This is because I_{ON} is proportional to the area of overlap between the gate and the source, as noted above. If the source edge is smooth, then gate LER results in variation in this overlap area; in contrast, if the source edge is perfectly correlated with the rough gate edge, then there will be no variation in this overlap area. The BTBT contour plots in Fig. 5.16 confirm that the tunneling area varies for the Smooth source edge case whereas it does not for the Rough source edge case.

The small variation in I_{ON} observed for the Rough source edge case is caused by the variation in L_{eff} which results in variation in V_T. As can be seen from the scatter plot in Fig. 5.17, there is still some weak correlation between I_{ON} and V_T for the Rough source edge case.

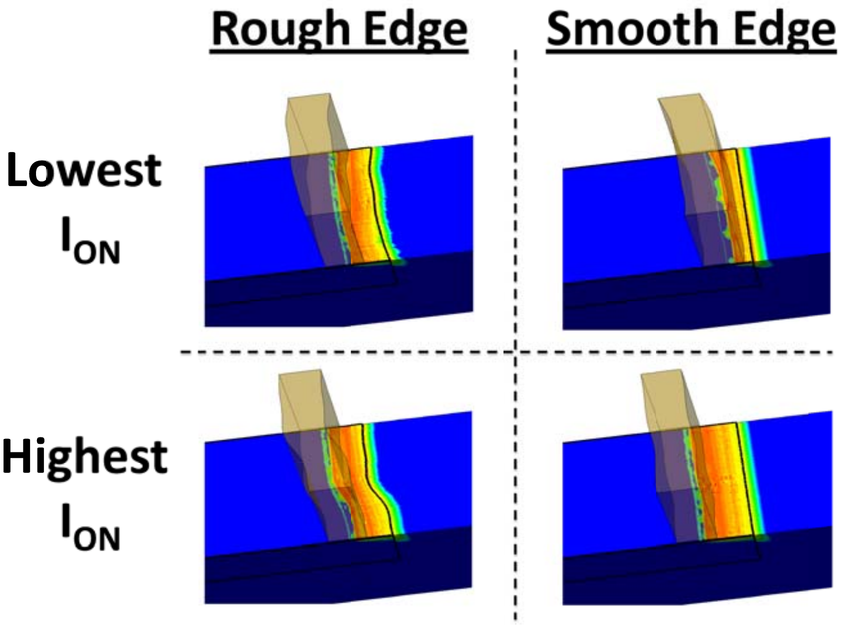


Fig. 5.16 Band-to-band tunneling contour plots for the TFET structures with the lowest and highest values of I_{ON}, for each of the Rough edge and Smooth edge LER cases. V_{GS} = 0.5 V and V_{DS} = 0.5 V. The gate electrode is not shown for clarity.

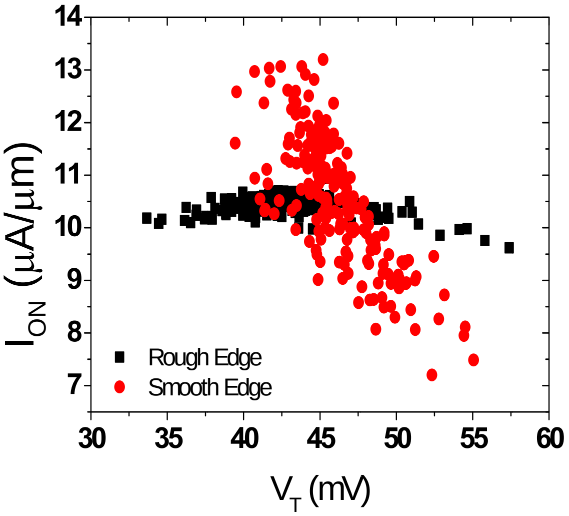


Fig. 5.17 Scatter plot of I_{ON} *vs.* V_T. I_{ON} is the drain current at $V_{GS} = V_{DS} = 0.5$ V, and V_T is the gate voltage corresponding to a drain current of 10^{-9} A/um with $V_{DS} = 0.5$ V. The Rough source edge case shows significantly smaller I_{ON} variation compared to the Smooth source edge case.

The impact of LER-induced variation on the trade-off between I_{ON} and I_{OFF} is shown in Fig. 5.18. (I_{ON} is taken to be the drain current at $V_{GS} = V_{DS} = 0.5$ V, and I_{OFF} is taken to be the drain current at $V_{GS} = 0$ V, $V_{DS} = 0.5$ V.) In this plot, this trade-off is adjusted by tuning the gate work function from 4.0 to 4.35 eV. Average values are indicated with open symbols. Relying on the fact that I_{ON} and $\log(I_{OFF})$ distributions are approximately Gaussian, the impact of 3σ variation to reduce I_{ON} and to increase $\log(I_{OFF})$ is indicated with filled symbols, and is seen to be significantly worse for the case of a Smooth source edge, primarily due to the larger variation in on-state current.

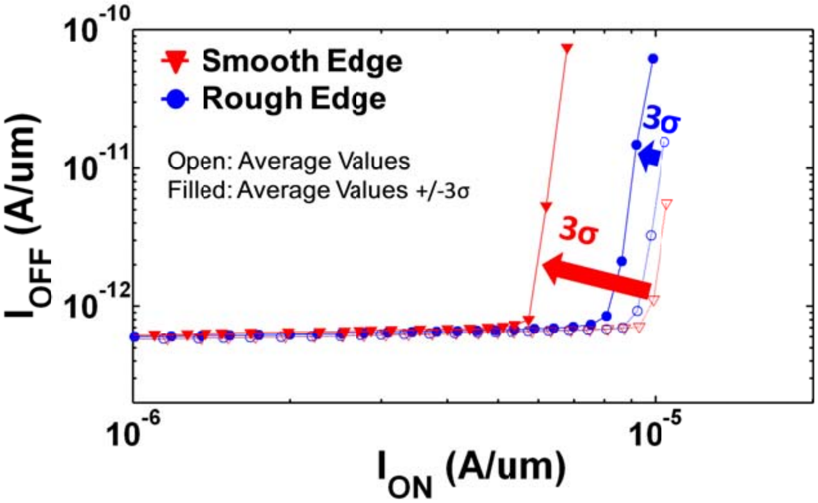


Fig. 5.18 Impact of LER-induced variation on the trade-off between I_{ON} and I_{OFF}, for the Ge-source TFET.

5.3.3 LER and RDF Co-Implementation

Atomistic doping was implemented using the Sano algorithm [17], together with gate LER, to assess the combined effect of these two random variation sources and elucidate any interaction effects. The assumptions and limitations of this RDF modeling approach as applied to TFETs are discussed in detail in [26]. Fig. 5.19 shows an example of a simulated device structure with both RDF and LER, Smooth edge case.

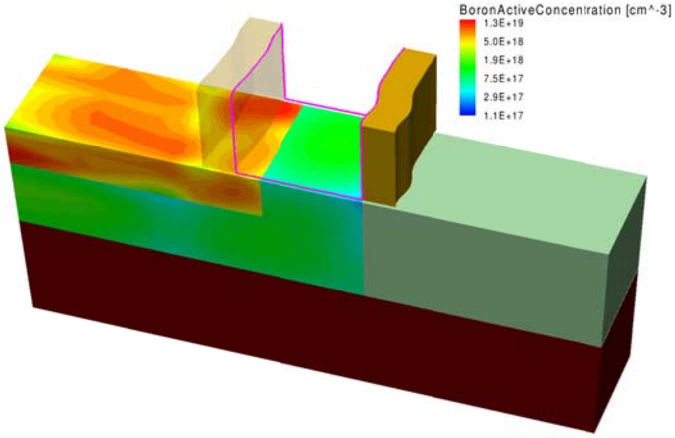


Fig. 5.19 Simulated 3D TFET structure with atomistic doping in the source and channel regions. The gate LER profile is reflected in the outer edge of the gate-sidewall spacer but not in the source edge profile. The gate electrode is not shown for clarity.

5.3.4 Impact of LER and RDF Together on TFET Performance

Atomistic doping within the source, channel, and drain regions is applied to the 200 individual TFET structures with gate LER (Smooth source edge case). The simulation is computationally expensive even with an optimized meshing strategy. One device simulation sweep (i.e. I_{DS}-V_{GS}) can take about 30-40 hours to complete running on 4 parallel 2.4 GHz CPU cores.

The ensembles of simulated transfer characteristics are shown in Fig. 5.20. Fig. 5.21 plots the corresponding V_T distribution, which is approximately Gaussian with a value of standard deviation equal to 14 mV. In a previous study, σV_T due to RDF-induced variation was found to be 13.99 mV [26]. Thus, the additional effect of LER is almost negligible, *i.e.* most of the V_T variation is caused by RDF.

Fig. 5.22 shows the distributions of I_{ON} and log(I_{OFF}) for two values of gate work function. With a low gate work function (4.0 eV), I_{OFF} is determined by (weak) BTBT so that it is very sensitive to variations in the local electric field and hence RDF. To minimize I_{OFF} variation, a larger gate work function (4.1 eV) should be used so that I_{OFF} is determined by the leakage current of the reverse-biased p-n drain-source junction, *i.e.* so that I_{OFF} reaches the leakage floor. Variation in the leakage floor current is due primarily to RDF, as alluded to previously.

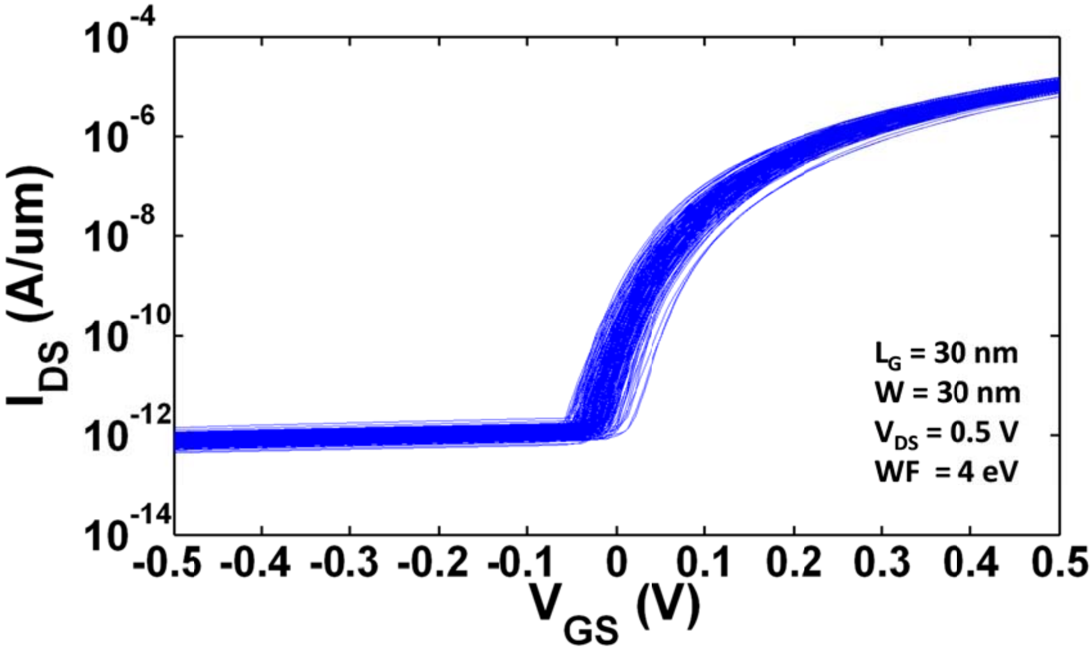


Fig. 5.20 Ensemble of simulated I_{DS}-V_{GS} characteristics for 200 planar Ge-source TFETs with gate LER (Smooth edge case) and RDF. V_{DS} = 0.5 V. The gate work function is set to 4.0 eV.

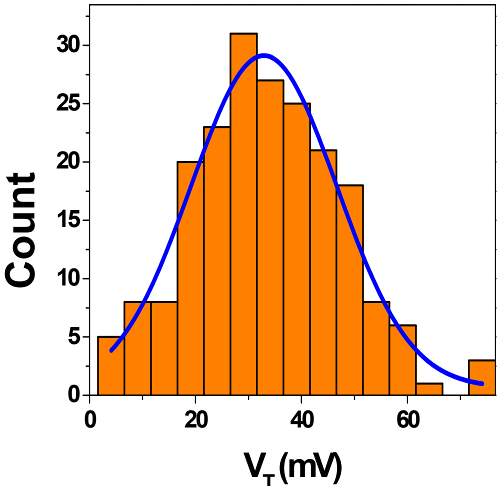


Fig. 5.21 Threshold voltage V_T distribution for the 200 simulated devices corresponding to Fig. 5.19. The blue line represents best-fit Gaussian curve.

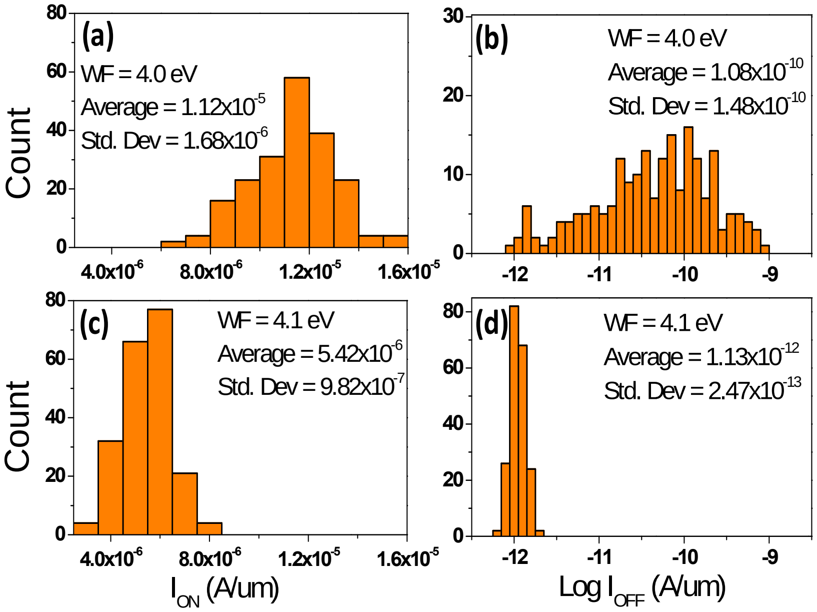


Fig. 5.22 Ge-source TFET I_{ON} and $\log(I_{OFF})$ distributions: (a)-(b) gate work function = 4.0 eV; (c)-(d) gate work function = 4.1 eV.

The impact of LER- and RDF-induced variations on the trade-off between I_{ON} and I_{OFF} is shown in Fig. 5.23. For reference, the curves showing the impacts of LER only (Smooth edge case) and RDF only are also included. The additional impact of LER can be seen to be a modest decrease in I_{ON} for $I_{OFF} > 2\times10^{-12}$ A/um. (For the Rough source edge case, the degradation in I_{ON} will be considerably less.)

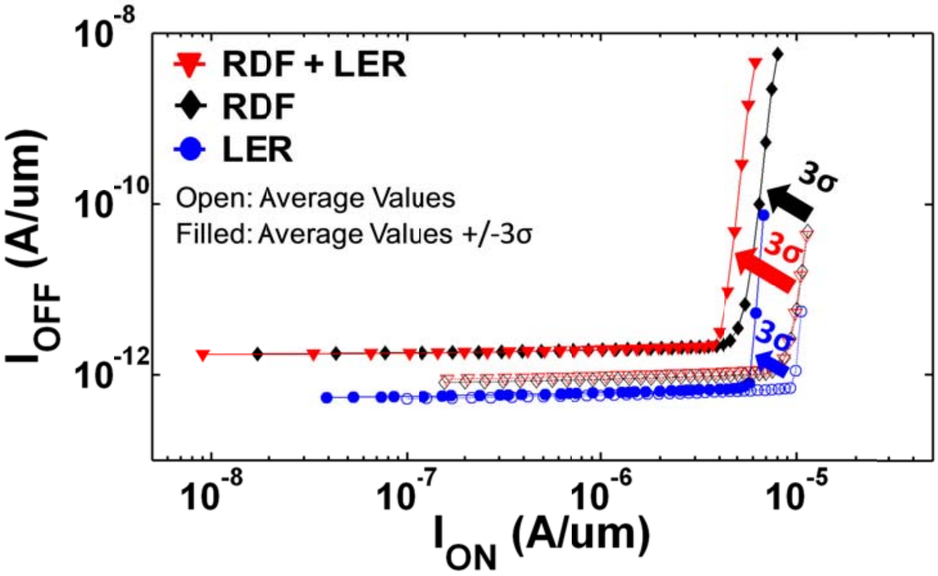


Fig 5.23 Impact of LER- and RDF-induced variations on the trade-off between I_{ON} and I_{OFF}, for the planar Ge-source TFET.

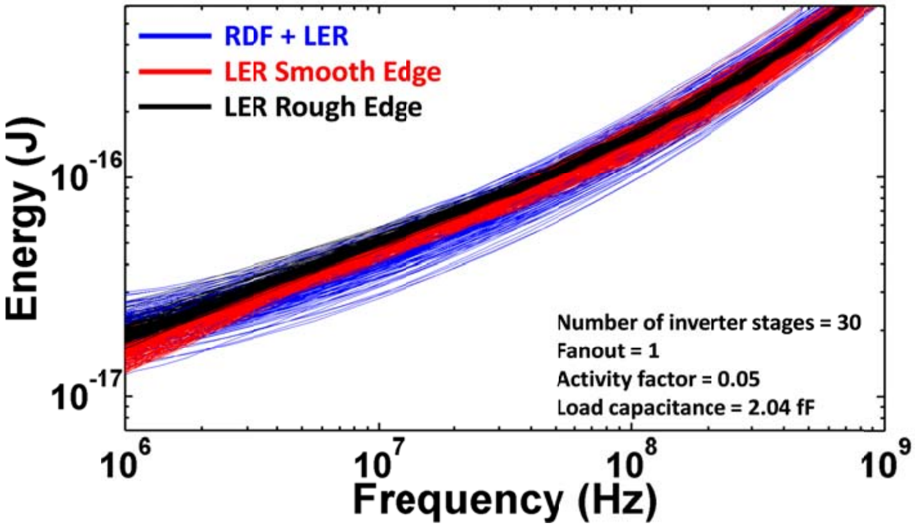


Fig. 5.24 Ensembles of energy vs. frequency characteristics for planar Ge-source TFET technology with line-edge roughness and/or random dopant fluctuation.

The impact of LER- and RDF-induced variations on planar Ge TFET circuit performance is shown in Fig. 5.24. Following the methodology described in [27-28], a chain of inverters (30 stages, fanout = 1, activity factor = 0.05, load capacitance = 2.04 fF = $2C_{gg}$) is used to assess the energy *vs.* frequency trade-off. The gate work function and the operating voltage V_{DD} are co-optimized to achieve the lowest energy per operation for a given frequency. It can be seen that the average energy per operation is higher for the case of LER with Rough source edge than for the case of LER with Smooth source edge, especially in the lower frequency range. This is because the Rough source edge case has slightly worse effective subthreshold swing (so that for a given I_{ON} required to achieve a particular frequency, I_{OFF} is higher). However, variability is smaller for the case of LER with Rough source edge because the gate-to-source overlap area is constant. RDF results in significantly increased variation, especially at frequencies below 100 MHz.

5.3.5 V_T Variation Comparison with the MOSFET

The individual contributions of RDF and LER to σV_T for the Ge-source TFET are compared against those reported for a planar bulk MOSFET of the same nominal gate length (30 nm) and channel width (~30 nm) [8], in Fig. 5.25. RDF-induced V_T variation for the Ge-source TFET is 62% lower than that for a MOSFET. Because the TFET is designed for tunneling to occur primarily vertically within the Ge source region [26], the impact of the lateral electric field (which varies with L_{eff}) is lower than that in a MOSFET which is designed for thermionic emission to occur primarily laterally from the source region to the channel region [12,20]. Therefore, LER-induced V_T variation for the TFET is much lower (by 87%) than that for the MOSFET. Additionally, the impact of gate LER induced threshold voltage variation in MOSFET is seen to be more than 50% of the RDF induced variation, while for TFET it accounts for less than 30%.

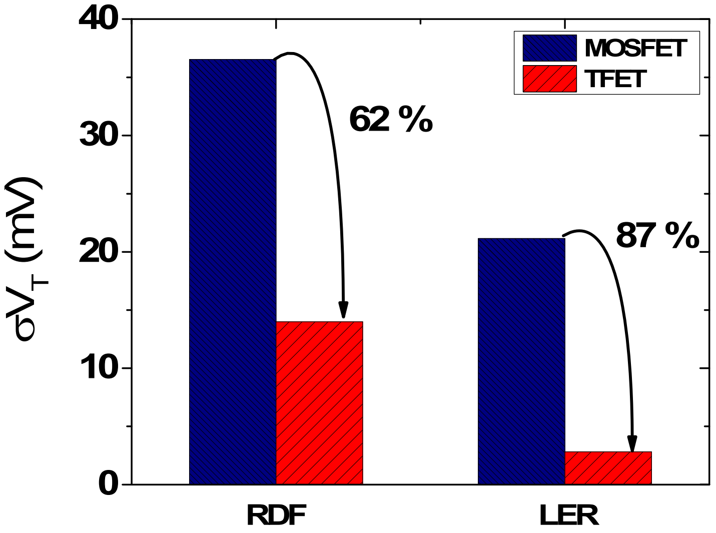


Fig. 5.25 Comparison of σV_T induced by RDF or LER in a planar bulk MOSFET *vs.* a planar Ge-source TFET.

5.4 Random Dopant Fluctuation Induced Variability in the Raised-Ge-Source TFET

5.4.1 Introduction

One of the main challenges for a tunnel FET is its relatively low on-state drive current compared to a MOSFET. This challenge can be overcome by using a smaller band-gap material such as InGaAs, SiGe, or Ge to help reduce the effective tunneling band gap [29-31,7]. However, the reported I_{ON} values from these works for operating voltage of 0.5V are approximately 1uA/um or less. In addition to band-gap engineering, the performance of a tunnel FET can be improved through optimization of the device geometry. Most tunnel FETs are designed for lateral/point tunneling. This approach can indeed achieve SS steeper than 60 mV/decade at room temperature. However, the major drawback is that the carrier injection happens across a very small area, and hence I_{ON} will also be small.

An alternative design to a lateral tunneling device is the vertical tunneling device where the tunneling happens within the source-region, perpendicular to the gate/gate-dielectric interface. A planar Ge-source TFET discussed previously is an example of a vertical tunneling device, which results in an improvement in I_{ON}. But as the device is turning on, there still exists a lateral tunneling component before the vertical tunneling component dominates at high gate voltage. This lateral tunneling ultimately degrades the overall subthreshold swing. To mitigate this issue, a raised-Ge-source TFET was proposed [14], wherein the source-region is fully elevated. The raised source design allows tunneling to happen almost entirely vertically, while virtually eliminating the lateral tunneling component. Moreover, the source region at which tunneling happens is electrically de-coupled from the drain field, which results in a smaller modulation of the tunneling barrier by the drain bias.

Variability in transistor performance worsens significantly as the gate length (L_g) is scaled down below 25 nm [8], and can limit reductions in supply voltage (V_{DD}) and hence power consumption, due to the need for design margin to achieve high yield [32-35]. Sources of random variation in TFETs include random dopant fluctuations (RDF), gate line-edge roughness (LER), and surface roughness [8],[36],[37]. With its improved electrostatic and higher drive current capability, a raised-Ge-source TFET is a promising device candidate for a low voltage operation. Nevertheless, one must carefully assess the performance degradation in the presence of variability. In the previous section, it was found that RDF is the dominant source of V_T variation for the planar Ge-source TFET [37]. In this section, the impact of RDF is investigated via TCAD simulation for the raised-Ge-source TFET design.

5.4.2 Device Structure and Simulation Approach

A cross-sectional view of the simulated n-channel TFET structure is shown in Fig. 5.26 (a). The nominal structure is optimized to achieve maximum I_{ON}/I_{OFF} for a supply voltage of 0.5V, similarly as in [34]. The nominal design parameter values are summarized in Table 5.3. Three-dimensional (3-D) device simulations are performed using Sentaurus [16]. Drift and diffusion models are used for carrier transport. (Due to the low operating voltage, ballistic transport and velocity overshoot are ignored.) A dynamic non-local tunneling model is used to simulate the tunneling process using calibrated tunneling coefficients A = 1.46×10^{17} $cm^{-3}\cdot s^{-1}$ and B = 3.59×10^{6} $V\cdot cm^{-1}$ [11,34]. The Modified Local Density Approximation (MLDA) captures quantum confinement effects, and the Oldslotboom model accounts for band-gap narrowing. A fixed charge density (8×10^{12} q/cm^2) at the Ge-source and gate oxide interface is assumed, based on previous fitting to experimental data [11,34].

To model RDF, device structures with atomistic doping profiles are generated following Sano's algorithm [17]. The algorithm accounts for the long-range Coulombic potential of the ionized dopant atoms, while the short-range potential is ignored to avoid unrealistic charge trapping and non-convergence issues. The short-range Coulombic potential is accounted for by the drift-diffusion simulator itself through appropriate mobility models. Fig 5.26 (b) shows an example of a device with randomly placed boron atoms within the source and body regions. An ensemble of 200 TFETs, each with uniquely randomized doping profiles, is simulated to assess the impact of RDF.

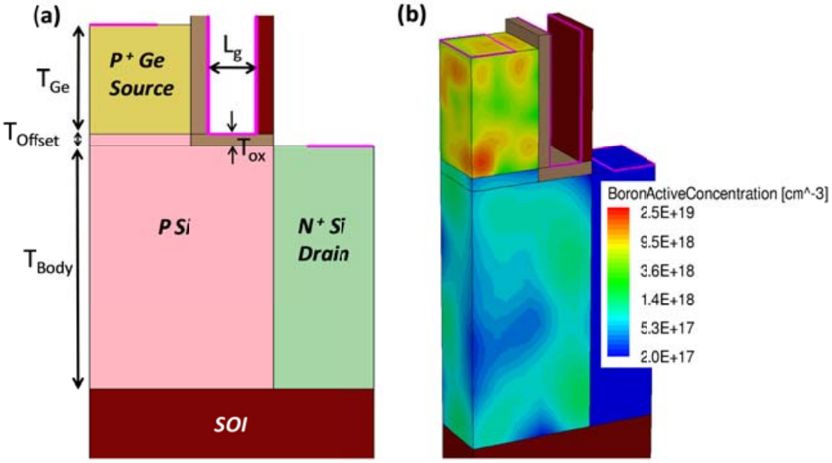


Fig 5.26 a) 2-D cross-section of the simulated raised Ge-source n-channel TFET b) An exemplary TFET structure with randomly placed boron atoms in the source and body regions.

TABLE 5.3
NOMINAL VALUES FOR RAISED-GE-SOURCE TFET DESIGN PARAMETERS

Device Parameter	Nominal Value
Gate Length (L_g)	14 nm
Oxide Thickness (T_{OX})	5.1 nm
Dielectric Relative Permittivity (ε_r)	23.4
Silicon Offset (T_{Offset})	5 nm
Body Thickness (T_{Body})	100 nm
Germanium Source Thickness (T_{Ge})	45 nm
Width (W)	30 nm
Source Doping (N_{Source})	10^{19} cm^{-3}
Body Doping (N_{Body})	10^{18} cm^{-3}
Drain Doping (N_{Drain})	10^{19} cm^{-3}
Gate work function	4 eV

5.4.3 Impact of RDF on DC Performance

Simulated I_{DS} *vs.* V_{GS} characteristics with $V_{DS} = V_{DD} = 0.5$V are shown in Fig. 5.27, for two cases: RDF only within the Ge source region (red curves) *vs.* RDF within all regions (blue curves). For reference, the simulated curve for the case of continuum doping profiles is also shown. It can be seen that the variation in $V_{T,Sat}$ (defined as the value of V_{GS} at I_{DS} = 1 nA/um with V_{DS} = 0.5V) is almost identical for the two RDF cases, with less than 2% difference in standard deviation ($\sigma V_{T,Sat}$). This indicates that V_T variation is due primarily to RDF within the source region, for the raised-Ge-source TFET. Since band-to-band tunneling occurs primarily within the source region in a raised-Ge-source TFET, variations in local source dopant concentration are expected to result in significant V_T variation. I_{ON} variation is due primarily to RDF within the body and drain regions which give rise to variations in parasitic series resistance.

It also can be seen from a comparison of the ensemble of curves with the nominal curve in Fig. 5.26 that RDF within the source region results in degraded SS and lower turn-on voltage. This is because BTBT no longer occurs uniformly along a single vertical plane, due to variations in local dopant concentration within the source, *i.e.* tunneling turns on/off at slightly different gate voltages [38]. Specifically, at locations where the effective dopant concentration is lower, BTBT turns on/off at lower V_{GS}. Note that the average gate voltage for which all BTBT paths (including those corresponding to the locations of highest effective dopant concentration) are activated is unchanged, so that V_T (defined at a lower level of current) is always lower than for the nominal case.

One of the advantages of the raised source design over the planar source design is that the influence of the drain voltage is reduced. As a result, Drain-Induced Barrier Tunneling (DIBT) – the difference between $V_{T,Sat}$ and $V_{T,Lin}$ (defined as the value of V_{GS} at I_{DS} = 1 nA/um with V_{DS} = 50 mV) – is much smaller for the raised source design. As indicated in Fig. 5.28, variation in DIBT is also significantly reduced, with ~ 35% lower standard deviation (σDIBT). The lower DIBT for the raised source design helps explain its relatively small σV_T (15.19 mV) as compared to that for a planar-Ge-source TFET with longer channel length ($\sigma V_{T,Sat}$ = 14mV) [26,37].

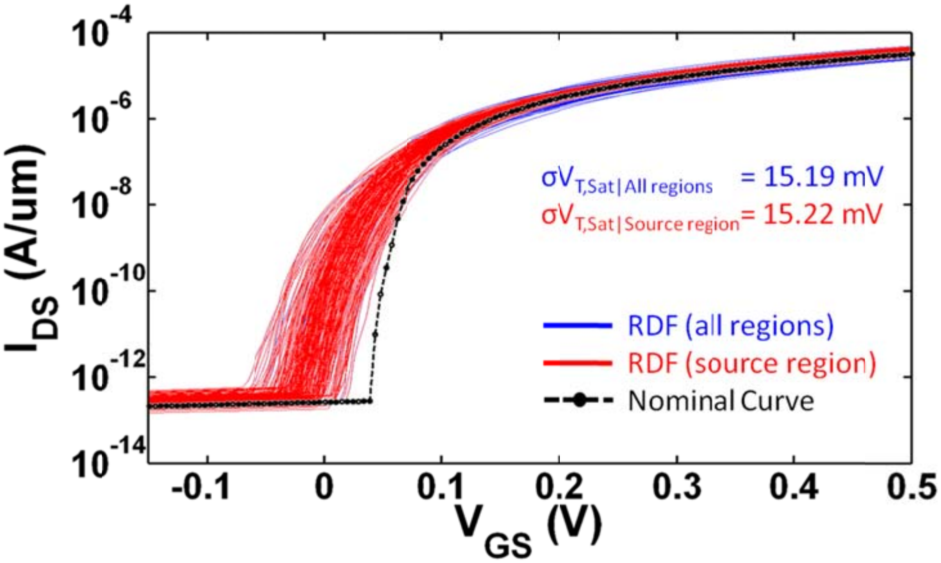


Fig 5.27 Ensembles of simulated I_{DS}-V_{GS} curves with V_{DS} = V_{DD} = 0.5 V. The simulated curve for the nominal TFET design with continuum doping profiles is shown in black.

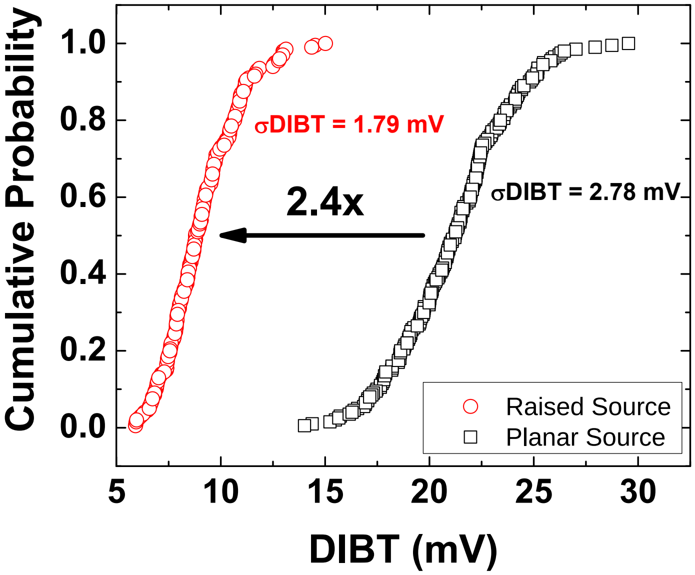


Fig 5.28 Cumulative probability plots of Drain-Induced Barrier Tunneling (DIBT) for raised-Ge-source (Red) *vs.* planar-Ge-source (Black) TFET designs. DIBT ≡ $V_{T,Lin} - V_{T,Sat}$.

5.4.4 Energy vs. Frequency Performance

Fig. 5.29 (a) plots the average value of I_{OFF} (defined as the value of I_{DS} for $V_{GS} = 0$ V and $V_{DS} = V_{DD} = 0.5$V) *vs.* the average value of I_{ON} (defined as the value of I_{DS} for $V_{GS} = V_{DS} = V_{DD} = 0.5$V), when the gate work function is varied from 4.35 eV down to 4.0 eV. In the regime where I_{OFF} is dominated by reverse-bias p-n junction leakage (*i.e.* when the onset of BTBT occurs at $V_{GS} > 0$V), there is minimal trade-off between increasing I_{ON} and increasing I_{OFF}. However, in the regime where I_{OFF} is dominated by BTBT (where $I_{ON} > 30$ uA/um), there is a large trade-off between increasing I_{ON} and increasing I_{OFF}. The boundary between these two regimes occurs at lower I_{ON}, ~10 uA/um, if RDF-induced variability (3σ increase in log(I_{OFF}) and 3σ decrease in I_{ON}) is taken into account.

The minimum energy per cycle *vs.* frequency for a 30-stage inverter chain (fanout = 1, activity factor = 0.05, load capacitance = 4.12 f F/um = 2× C_{gg}) is evaluated using the methodology described in [21]. Fig. 5.29 (b) compares the results of this analysis for the nominal raised-Ge-source TFET design, an ideal short-channel MOSFET design (SS = 60 mV/dec), and the ensemble of raised-Ge-source TFETs with atomistic doping profiles. The $I_{DS} - V_{GS}$ characteristics for the nominal TFET and ideal MOSFET are shown in the inset of Fig 5.29 (b).

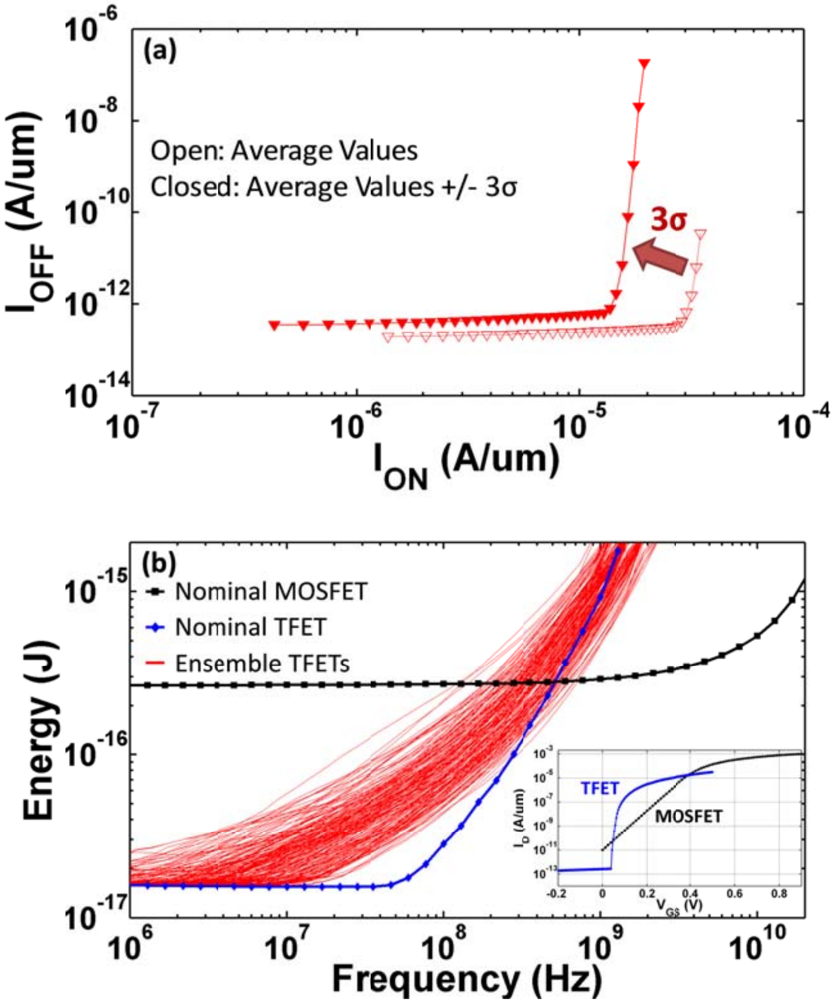


Fig 5.29 a) I_{ON}-I_{OFF} trade-off with 3σ variability taken into account b) Energy vs. frequency analysis. Ensemble TFET curves are shown in red; Nominal TFET is shown in black; Nominal "ideal" MOSFET is shown in blue.

RDF-induced degradation in SS results in higher energy per operation (due to larger I_{OFF}) for frequencies below 300 MHz. At higher frequencies, RDF-induced variation in I_{ON} becomes more significant, so that the performance of the nominal TFET design lies approximately in the middle of the ensemble curves. From a comparison of the curves for the nominal TFET and ideal MOSFET designs, the TFET is projected to provide for lower energy operation at frequencies up to 500 MHz, due to its smaller effective SS and lower I_{OFF}.

5.5 Summary

The impact of random dopant fluctuations (RDF) on the performance of optimally designed Ge-source TFETs is relatively modest (σV_{TH} < 20 mV at L_g = 30 nm) compared to a MOSFET of similar gate length (σV_{TH} ~ 36 mV) [8]. RDF in the source region is found to be the main contributor to σV_{TH} for the vertical tunneling TFET design (moderately doped source).

RDF in the source and channel regions are equally important for the lateral tunneling TFET design (heavily doped source). Analysis of variations in I_{ON} and I_{OFF} are necessary to assess the impact of RDF on energy-delay performance, because V_{TH} is not a good proxy for these (*i.e.* σV_{TH} does not fully capture the effect of variability on device performance). Accounting for the impact of RDF-induced variations, the vertical tunneling TFET design is best for I_{ON} in the range >3uA/um. Unlike MOSFET, there is a weak dependence on TFET σV_{TH} and the drain to source voltage V_{DS}.

The impact of gate LER on the performance and variability of the planar Ge-source TFET is relatively small as compared to the impact of RDF, and it is minimized if the source edge is perfectly correlated to the gate edge. The impact of LER-induced variation in I_{ON} becomes significant for applications requiring transistor drive current above ~5 uA/um for V_{DD} = 0.5 V. From an energy-performance point of view, the impact of LER on variation in minimum operating energy becomes comparable to that of RDF at frequencies greater than 100 MHz.

RDF within the source region results in degraded SS and lower turn-on voltage for the raised-Ge-source TFET design. However, drain-induced barrier tunneling is further mitigated with the raised source design, resulting in a relatively small threshold voltage variation, as compared to the planar-Ge-source design with a longer channel length. For these reasons, the raised-Ge-source TFET is a promising candidate for low-power digital logic applications requiring operating frequencies below 500 MHz.

5.6 References

[1] S. Hanson, B. Zhai, K. Bernstein, D. Blaauw, A. Bryant, L. Chang, K. K. Das, W. Haensch, E. J. Nowak, and D. M. Sylvester, “Ultralow-voltage, minimum-energy CMOS,” *IBM Journal of Research and Development*, vol. 50, no. 4/5, pp. 469-490, 2006.

[2] W. Reddick and G. Amaratunga, “Silicon surface tunnel transistor,” *Appl.Phys. Lett.*, vol. 67, no. 4, pp. 494–496, Jul. 1995.

[3] J. Appenzeller, Y.-M. Lin, J. Knoch, and Ph. Avouris, “Band-to-Band Tunneling in Carbon Nanotube Field-Effect Transistors,” *Phys. Rev. Lett.,* 93(19), pp. 196805-1-4, 2004.

[4] K. Bhuwalka, J. Schulze, and I. Eisele, “Performance Enhancement of Vertical Tunnel Field-Effect Transistor with SiGe in the δp+ Layer,” *Jap. J. of Appl. Phys.*, 43(7A), pp. 4073-4078, 2004.

[5] Q. Zhang, W. Zhao, and A. C. Seabaugh, “Low-subthreshold-swing transistors,” *IEEE Electron Dev. Lett.,* vol. 27, pp. 297-300, 2006.

[6] W. Y. Choi, B.-G. Park, J. D. Lee, and T.-J. K. Liu, “Tunneling field-effect transistors (TFETs) with subthreshold swing (SS) less than 60 mV/dec,” *IEEE Electron Device Lett.*, vol. 28, no. 8, pp. 743–745, Aug. 2007.

[7] S. H. Kim, H. Kam, C. Hu, and T.-J. K. Liu, “Ge-source tunnel field effect transistors with record high ION/IOFF,” in *VLSI Symp. Tech. Dig.*, 2009, pp. 178–179.

[8] A. Asenov, “Simulation of statistical variability in nano MOSFETs,” *IEEE Symp. VLSI Technol., Dig. Tech. Papers*, Jun. 2007, pp. 86–87.

[9] A. Asenov, A. Brown, J. Davies, S. Kaya, and G. Slavcheva, “Simulation of intrinsic parameter fluctuation in decananometer and nanometer-scale MOSFETs”, *IEEE Trans. on Electron Devices*, vol. 50, 2003, pp.1837-1852.

[10] K. Boucart, W. Riess, and A. M. Ionescu, “A Simulation-based Study of Sensitivity to Parameter Fluctuations of Silicon Tunnel FETs,” *ESSDERC*, September 2010, pp. 345-348.

[11] S. H. Kim, Z. A. Jacobson, and T.-J. K. Liu, “Impact of body doping and thickness on the performance of Germanium-source TFETs,” *IEEE Trans. Electron Devices*, vol. 57, no. 7, Jul. 2010, pp. 1710–1713.

[12] K. Boucart and A. M. Ionescu, “Length scaling of the Double Gate Tunnel FET with a high-K gate dielectric,” Solid-State Electronics, vol. 51, issue 11-12, Nov-Dec 2007, pp. 1500-1507.

[13] Sentaurus User’s Manual, Synopsys, Inc., Mountain View, CA, 2010.

[14] S.H. Kim, S. Agarwal, P. Matheu, C. Hu, and T.-J. K. Liu, “Tunnel Field Effect Transistor With Raised Germanium Source,” *IEEE Electron Device Letters*, vol. 31, no. 10, Oct. 2010, pp 1107-1109.

[15] Z. A. Jacobson, S. Jin, S. H. Kim, P. Matheu, and T.-J. K. Liu, “Tunneling Model Calibration and Source Design Optimization for a Planar Ge-Source TFET,” *IEEE Trans. Electron Devices (accepted), 2011.*

[16] D. Kuzum, A. J. Pethe, T. Krishnamohan, and K. C. Saraswat, “Ge (100) and (111) N- and P-FETs with high mobility and low-T mobility characterization,” *IEEE Trans. Electron Devices*, vol. 56, no. 4, Apr. 2009, pp. 648–655.

[17] N. Sano, K. Matsuzawa, M. Mukai, and N. Nakayama, “Role of long-range and short-range coulomb potentials in threshold characteristics under discrete dopants in sub-0.1 μm Si-MOSFETs,” in *IEDM Tech. Dig.*, 2000, pp. 275-278.

[18] C. Shin, X. Sun, and T.-J. King Liu, "Study of random-dopant-fluctuation (RDF) effects for the trigate bulk MOSFET," *IEEE Trans. Electron Devices*, vol.56, no. 7, Jul. 2009, pp.1538-1542.

[19] K. Boucart and A. M. Ionescu, "A new definition of threshold voltage in tunnel FETs," *Solid State Electron*, vol. 92, no. 9, Sep. 2008, pp. 1318–1323.

[20] W. G. Vandenberghe, A. S. Verhulst, G. Groeseneken, B. Soree, and W. Magnus, "Analytical Model for Point and Line Tunneling in a Tunnel Field-Effect Transistor," in *Proc. SISPAD*, Sep. 2008, pp. 137-140.

[21] H. Kam, "MOSFET Replacement Devices for Energy-Efficient Digital Integrated Circuits," Ph.D. dissertation, Univ. California, Berkeley, CA, May 2008.

[22] M.J.M. Pelgrom, H. P. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications," in *IEDM Tech. Dig.,* 1998, pp. 915.

[23] O. Faynot, F. Andrieu, O. Weber, C. Fenouillet-Béranger, P. Perreau, J. Mazurier, T. Benoist, O. Rozeau, T. Poiroux, M. Vinet, L. Grenouillet, J.-P. Noel, N. Posseme, S. Barnola, F. Martin, C. Lapeyre, M. Cassé, X. Garros, M.-A. Jaud, O. Thomas, G. Cibrario, L. Tosti, L. Brévard, C. Tabone, P. Gaud, S. Barraud, T. Ernst and S. Deleonibus, "Planar Fully Depleted SOI Technology: a powerful architecture for the 20nm node and beyond," in *IEDM Tech. Dig.,* 2010, pp.51-53.

[24] J. A. Croon, G. Storms, S. Winkelmeier, I. Pollentier, M. Ercken, S. Decoutere, W. Sansen, H.E. Maes, "Line edge roughness: characterization, modeling and impact on device behavior," in *IEDM Tech. Dig.,* pp. 307-310, 2002.

[25] A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," *IEEE Trans. Electron Devices*, vol 50, no. 5, pp. 1254-1260, May 2003.

[26] N. Damrongplasit, C. Shin, S. H. Kim, R. Vega, and T. –J. K. Liu, "Study of Random Dopant Fluctuation Effects in Germanium-Source Tunnel FETs," *IEEE Trans. Electron Devices*, vol. 58, no. 10, Oct. 2011, pp. 3541-3548.

[27] H. Kam, T.-J. K. Liu, E. Alon, M. Horowitz, "Circuit-Level Requirements for MOSFET-Replacement Devices," in *IEDM Tech. Dig.*, pp. 427, 2008.

[28] H. Kam, T.-J. K. Liu, and E. Alon, "Design requirements for steeply switching logic devices," *IEEE Trans. Electron Devices*, vol 59, no. 2, Feb. 2012, pp. 326-334.

[29] A. M. Ionescu and H. Riel, “Tunnel field-effect transistors as energy-efficient electronic switches” *Nature* 479(7373): 329-337, Nov 2011.

[30] G. Dewey; B. Chu-Kung; J. Boardman; J.M. Fastenau; J. Kavalieros; R. Kotlyar; W.K. Liu; D. Lubyshev; M. Metz; N. Mukherjee; P. Oakey; R. Pillarisetty; M. Radosavljevic; H.W. Then; R. Chau, R., “Fabrication, Characterization, and Physics of III-V Heterojunction Tunneling field Effect Transistors (H-TFET) for Steep Sub-Threshold Swing,” in *IEDM Tech. Dig.,* pp. 33.6.1-33.6.4, Dec. 2011.

[31] A. Villalon; C. Le Royer; M. Casse; D. Cooper; B. Previtali; C. Tabone; J. Hartmann; P. Perreau; P. Rivallin; J. Damlencourt; F. Allain; F. Andrieu; O. Weber; O. Faynot; T. Poiroux, “Strained tunnel FETs with record ION: first demonstration of ETSOI TFETs with SiGe channel and RSD,” in *VLSI Symp. Tech. Dig.,* June 2012, pp. 49-50.

[32] T. Hiramoto, “Ultra-low-voltage operation: Device perspective,” *ISLPED.* , Aug. 2011, pp. 59-60.

[33] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, K. Bernstein, “ Scaling, power, and the future of CMOS,” in *IEDM Tech. Dig*., pp. 7–15, 2005.

[34] B. H. Calhoun, A. Wang and A. Chandrakasan, “Modeling and sizing for minimum energy operation in subthreshold circuits,” *IEEE Journal of Solid State Circuits*, Vol. 50, pp. 1778-1786, 2005.

[35] P. Zuber. M. Miranda, M. Bardon, S. Cosemans, P. Roussel, P. Dobrovolny, T. Chiarella, N. Horiguchi, A. Mercha, T.-Y. Hoffmann, D. Verkest, S. Biesemans, “Variability and technology aware SRAM Product yield maximization,” in *VLSI Symp. Tech. Dig.* pp.222,223, 14-16 June 2011.

[36] F. Conzatti, M. G. Pala, and D. Esseni, “Surface-Roughness-Induced Varaibility in Nanowire InAs Tunnel FETs,” *IEEE Electron Device Letters,* vol.33, no. 6, Jun. 2012, pp. 806-808.

[37] N. Damrongplasit, S. H. Kim, C. Shin, and T. –J. K. Liu, “, Impact of Gate Line-Edge Roughness (LER) Versus Random Dopant Fluctuations (RDF) on Germanium-Source Tunnel FET Performance,” *IEEE Trans. Nanotechnology*, vol. 12, issue. 6, Nov. 2013, pp. 1061-1067.

[38] C. Hu, P. Patel, A. Bowonder, K. Jeon, S. H. Kim, W. Y. Loh, C. Y. Kang, J. Oh, P. Majhi, A. Javey, T. –J. K. Liu and R. Jammy, “Prospect of tunneling green transistor for 0.1 V CMOS”, in *IEDM Tech. Dig*., pp. 387-390, 2010.

Chapter 6

Conclusion

6.1 Summary and Contributions of This Work

The planar bulk-Si MOSFET has been the workhorse transistor structure for CMOS integrated circuits over the last four decades. But with continued transistor scaling, its performance and variability worsen due to short channel effects, resulting in excessive OFF-state leakage current and increased reliability problems. Advanced transistor structures such as the planar FD-SOI MOSFET and 3D FinFET can suppress these undesirable effects, but at the same time they can substantially increase process complexity and manufacturing cost. Super steep retrograde (SSR) channel doping is shown to have the potential for extending the scalability of planar-bulk CMOS technology at minimal cost, which is ideal for mobile applications which are cost-sensitive. At the 28nm technology node, a SSR channel doping profile can reduce σV_T variability by 40 – 50 % compared to uniform doping profile. Estimation of 6-T SRAM cell yield (6σ) shows a 33% reduction in V_{MIN}, which provides for more than 50% dynamic power savings.

Traditional Monte Carlo SPICE models are inadequate for explaining and predicting variability in device performance at advanced technology nodes. Accurate modeling of V_T and DIBL variability is essential for estimating SRAM cell yield and for estimating r_{out} variability in analog devices. To meet this need, a physically-based variability model is developed to explain variations in V_T and DIBL, as well as the correlation between them. The model shows an excellent match to measured data for a 32nm HKMG transistor technology in production for both SRAM and analog devices. It analyzes the forward-and reverse-mode characteristics of MOSFETs. Four fundamental variability components are proposed including Chip Mean, Symmetric, Asymmetric, and L_{EFF} variability. The results indicate that random asymmetric variation (e.g. random dopant fluctuations due to halo and/or source-drain extension doping) is responsible for the weak correlation of $V_{T,\,Lin}$ *vs.* $V_{T,\,Sat}$, the non-Gaussian distribution of DIBL, and the increased $\sigma V_{T,\,MM}$ with increased drain bias V_{DS}.

A test chip has been designed to investigate the sources of variability in transistor performance. A device characterization array comprising mismatch transistor pairs and transistors with different layout proximity is used to study random and systematic variability in planar bulk and FD-SOI MOSFETs in a 28nm technology. To study the impact of transistor variability on SRAM cell performance, a padded-out SRAM array implemented in a 16nm FD-SOI technology is designed to examine the individual transistor characteristics and to correlate them to SRAM cell performance metrics. Multiplexer and decoder circuits are necessary for accessing each device under test (DUT), given the limited number of I/O pads on a chip. Automated testing allows for the characterization of many devices without the need for constant supervision. Parasitic effects inherent to test chips such as array leakage current or voltage drop along the wires can be mitigated by strongly turning off the devices which are not under test and implementing the Kelvin Force/Sense measurement technique, respectively. To allow for accurate reliability testing, a switchbox can be used to quickly switch between regular I-V sweep mode and voltage stress mode, which is important for NBTI/PBTI and RTN characterizations that rely on short time-constant.

The tunnel FET has emerged as a promising candidate to replace the MOSFET for low-power applications due to its potential for achieving higher ON/OFF current ratio at low operating voltage. In particular, a planar Ge-Source TFET structure provides for better I_{ON} as compared to silicon TFET without incurring too much process complexity which comes with the use of exotic materials such as III-V. The impact of RDF on threshold voltage variation of a planar Ge-Source TFET is less than 20 mV at L_g = 30 nm (compared to 36 mV of a planar MOSFET of similar gate length) [1]. Depending on the relative doping between the source and the channel regions of a TFET, the tunneling mechanism can occur predominantly in either the lateral or the vertical direction. The effect of atomistic doping in the source region is found to be the main contributor to σV_T for a vertical tunneling TFET design. Atomistic doping effects in the source and channel regions are equally important for the lateral tunneling TFET design. Analysis of I_{ON} and I_{OFF} variation is required to accurately evaluate the impact of RDF on the energy-delay performance tradeoff for a TFET. The vertical tunneling TFET design shows lower energy per operation as compared to the lateral design for I_{ON} > 3 uA/um. Furthermore, the impact of gate LER on the source edge profile of a TFET is found to have minimal impact on V_T variation. The impact of LER on I_{ON} variation can be significant when drive current is above ~ 5uA/um for V_{DD} = 0.5V. To further improve the performance of a planar Ge-Source TFET, a raised source structure can be used. Drain induced barrier tunneling (DIBT) is found to be 2.4 times smaller and DIBL variation is also reduced by 35% for a raised-Ge-source TFET. From a comparison of energy-delay performance accounting for the effect of RDF, the raised-Ge-source TFET can provide energy savings for low-power applications at operating frequencies below 500 MHz.

6.2 Suggestions for Future Work

The use of SSR channel doping in the conventional planar bulk MOSFET design is advantageous for low-power and low-cost applications. The benefit of this design can be extended even further when body biasing is exploited at the circuit level [2]. Forward-biasing can be used to lower the V_T of the device, increasing its drive current. On the other hand, reverse-biasing can be used to raise the V_T of the device to reduce OFF-state leakage current. It would be informative to assess the improvement gain with this optimized biasing scheme for different circuit topologies.

To study V_T and DIBL variability, the physically based model developed in this work can be made more useful for digital and analog circuit designers if it can be incorporated into a SPICE model. By utilizing the four extracted fundamental variability components as the base parameters, a wrapper can be implemented around industry-standard compact models such as BSIM or PSP to better predict statistical variations in transistor performance.

A device characterization array enables a large number and variety of transistors to be characterized on a single chip,but the leakage current contributions from non-accessed devices within the array can significantly affect the accuracy of the measurement of the DUT characteristics as compared to direc probing of individual devices. Thus, better access circuit designs or array partitioning to help mitigate unwanted leakage current is desired. For dynamic characterization of trap-related phenomena that requires fast rise or fall times, a built-in self-test (BIST) circuit can be implemented to generate the desired waveform on the chip, rather than relying on an external pulse generator to supply signals through long wires. (By the time the signal finally arrives at the device, the shape of the waveform can be altered significantly.)

First principles calculations and atomistic simulation can provide for more accurate modeling of novel switching devices such as TFETs. However, due to computational complexity, these approaches are practically limited to small device sizes and numbers, and are not suitable for simulating a large number of transistors required for variability study. Device simulation using TCAD modeling is the preferred approach and offers a good trade-off between accuracy and computation time. Nevertheless, better calibration of the models (band-to-band tunneling, quantum confinement incluing dimensionality effects, etc.) to experimental results or atomistic simulations will be necessary in order to accurately predict TFET performance for different doping values and biasing conditions. More experimental TFET data is also desired to verify the accuracy of the models variability sources such as RDF and LER on TFET characteristic.

6.3 The Variability Challenge Ahead

The semiconductor industry has seen a remarkable progression in integrated circuit technology over the years, which helped to usher in the personal computing era, World Wide Web, mobile and cloud based computing, and the nascent market of Internet of Things (IoT). Underlying this success is the unwavering determination and resiliency to keep pace with Moore's Law, in spite of all the technical challenges. Unfortunately, all exponential growth trends eventually come to an end, and Moore's Law is no exception. There are signs that the end of Moore's Law may not be far out. Market analysis shows that the cost of manufacturing a transistor is no longer decreasing, starting from the 28 nm technology node as shown in Fig. 6.1 [3]. Since the motivation behind Moore's Law is to reduce cost, the recent trend for transistor manufacturing cost is of major concern. Chip design cost has also risen rapidly with each new technology node [3]. In fact, an apparent slowdown can be observed when looking at the volume production at a given technology node *vs.* the year it is introduced. It can be seen that starting from the 32 nm technology node, the introduction of new generation of chips is taking longer than before [4].

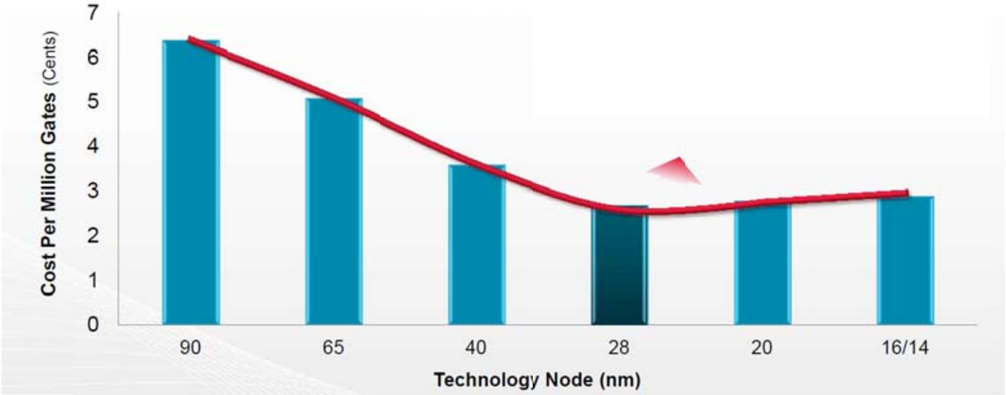


Fig. 6.1 Transistor manufacturing cost *vs.* technology node. According to the prediction, this cost stops decreasing after the 28nm technology node (adapted from [3]).

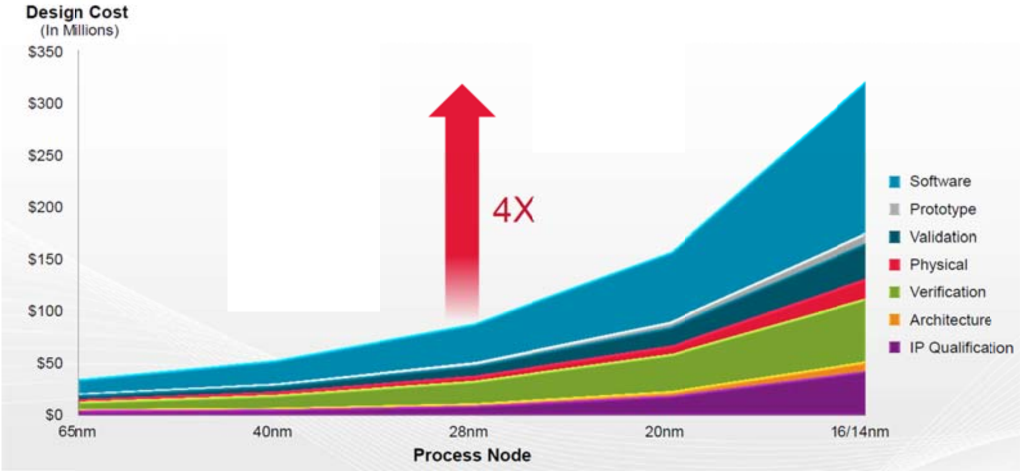


Fig. 6.2 Design cost *vs.* technology node (adapted from [3]).

One of the culprits contributing to the slowdown, which may even become the showstopper of Moore's Law, is variability in device performance, which ultimately lowers chip yield and thereby drives up cost. To mitigate this issue as transistors are further scaled into the sub-10 nanometer regime, co-optimization of device and circuit designs will be indispensible. For example, in an SRAM array, replacing the planar bulk MOSFET with advanced transistor structures such as the FD-SOI MOSFET or FinFET can improve cell stability and performance [5,6]. But the benefit stemming from device innovation alone is not enough. Alternative SRAM cell designs such as 8-T and 10-T, or circuit-assist techniques such as V_{DD} collapse, word-line boost, and negative bit-line will become more prevalent in future technology generations (Fig.6.3) [7-9].

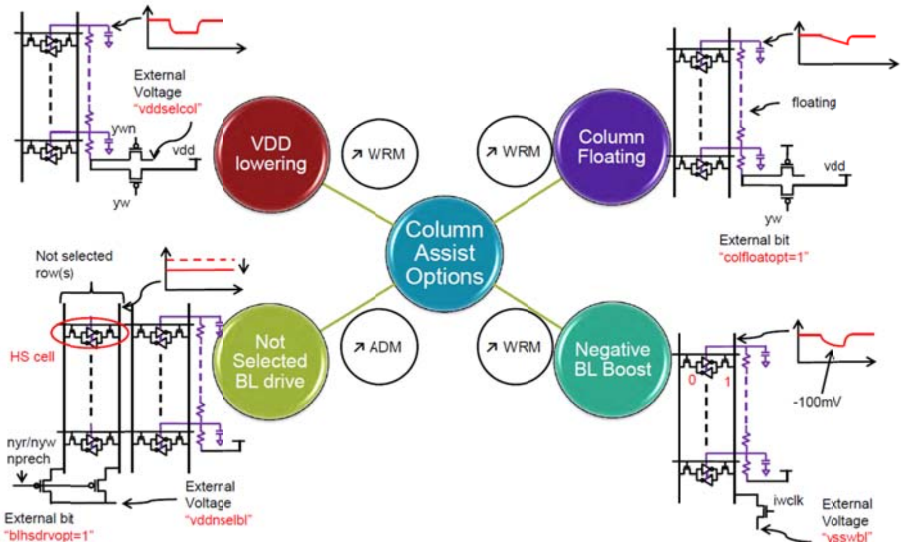


Fig. 6.3 Various circuit-assist techniques to help improve SRAM cell performance and stability [9].

Secondly, there has been a growing effort to search for a MOSFET-replacement device, one that can operate at higher frequency and/or with less energy. Even though this new device may have the potential to outperform the MOSFET, its performance benefits will be limited by variability that is unique to the device. For instance, a carbon nanotube transistor can suffer from V_T variability due to poor substrate passivation, a graphene transistor can be affected the edge roughness of the graphene nanoribbon, resulting in large device-to-device variability [10-11]. Therefore, it is imperative to have a good understanding of how variability will impact device performance, so that its effect can be accurately quantified and an appropriate guard band can be included in the circuit design to ensure proper circuit operation.

Lastly, if one were to consider variability to be an inextricable part of nanotechnology, then one might as well embrace it instead of trying to eliminate it. The mainstream computing paradigm relies on the Von Neumann digital system architecture in which distinct states are represented with sufficient noise margin between them. Since variability can reduce this noise margin, it is considered undesirable in a digital system. However, if one were to consider an alternative computing architecture, like the architecture used by the brain, variability is not only acceptable, but also preferred [12]. In fact, the neuron networks in the brain are considered to be very noisy, yet the brain is able perform complex tasks with minimal power consumption under such constraints [13].

There is no denying that the end of Moore's Law is inevitable. But even when physical scaling of transistor dimensions eventually runs out of steam, performance improvement and manufacturing cost reduction can still be achieved through innovations in device design, circuit design, interconnect design, packaging technology, and computer architecture. Continued investment in R&D by leading companies and government agencies will pave the way for further advancements in IC technology which will bring benefits to society for generations to come.

6.4 References

[1] A. Asenov, “Simulation of statistical variability in nano MOSFETs,” *IEEE Symp. VLSI Technol., Dig. Tech. Papers*, Jun. 2007, pp. 86–87.

[2] Suvolta, “Body Effect and Body Biasing,” *Technical Briefing*, June 2011.

[3] Broadcom, *Analyst day 2013*, Dec. 2013.

[4] C. G. Dieseldorff, C. Tseng, “Technology Node Transitions Slowing Below 32nm,” Industry Research & Statistics, SEMI.org, June 2014.

[5] T. Song; W. Rim; J. Jung; G. Yang; J. Park; S. Park; K.-H Baek; S. Baek; S.-K. Oh; J. Jung; S. Kim; G. Kim; J. Kim; Y. Lee; K. S. Kim; S.-P. Sim; J. S. Yoon; K.-M. Choi, “13.2 A 14nm FinFET 128Mb 6T SRAM with VMIN-enhancement techniques for low-power applications,” *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International* , vol., no., pp.232,233, 9-13 Feb. 2014.

[6] R. Ranica; N. Planes; O. Weber; O. Thomas; S. Haendler; D. Noblet; D. Croain; C. Gardin; F. Arnaud, “FDSOI process/design full solutions for ultra low leakage, high speed and low voltage SRAMs,” *VLSI Technology (VLSIT), 2013 Symposium on* , vol., no., pp.T210,T211, 11-13 June 2013.

[7] L. Chang; R.K. Montoye; Y. Nakamura; K.A. Batson; R.J. Eickemeyer; R.H. Dennard; W. Haensch; D. Jamsek, “An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches,” *Solid-State Circuits, IEEE Journal of* , vol.43, no.4, pp.956,963, April 2008.

[8] H. Noguchi; Y. Iguchi; H. Fujiwara; Y. Morita; K. Nii; H. Kawaguchi; M. Yoshimoto, “A 10T Non-Precharge Two-Port SRAM for 74% Power Reduction in Video Processing,” *VLSI, 2007. ISVLSI '07. IEEE Computer Society Annual Symposium on* , vol., no., pp.107,112, 9-11 March 2007.

[9] R. Aitken, “The Challenges of FinFET Design,” *Electronic Design Process Symposium (EDPS),* April 2013.

[10] A. D. Franklin, G. S. Tulevski, S.-J. Han, D. Shahrjerdi, Q. Cao, H.-Y. Chen, H.-S. P. Wong, W. Haensch, “Variability in carbon nanotube transistors: improving device-to-device consistency,” *ACS Nano* **6**, 1109–1115 (2012).

[11] Y. Yoon; J. Guo, “Effects of edge roughness in graphene nanoribbon transistors,” *Appl. Phys. Lett.* **91**, 073103 (2007).

[12] J. Rabaey, “Information Technology and Computational Neuroscience – What can they learn from each other?” *Distinguished Lecture Program, LLEUVEN Center on Information and Communication Technology (LICT)*, March 2013.

[13] A. A. Faisal, L. P. J. Selen, D. M. Wolpert, “Noise in the nervous system,” *Nat. Rev. Neurosci.*, 9 (2008), pp. 292–303